A Pipeline for Automated Analysis of Factual Claims in YouTube Transcripts

Farnoush Nilizadeh, Ali Ansari,
Sanobar Rustamova, Abhudaya Shrivastava

December 2025

Abstract

This report details the design, implementation, and rigorous validation of a five-phase Natural Language Processing (NLP) pipeline for the automated extraction and analysis of factual claims from unstructured YouTube video transcripts. The focus is on high-volume, noisy health discourse surrounding GLP-1 Agonists (e.g., Ozempic/Wegovy). The core engineering challenge was converting raw video data into a structured, queryable knowledge base of atomic claims, strategically counteracting the inherent inconsistency and data corruption issues common in social media scraping. We achieved high reliability on our core extraction tasks (F1 score of 0.94 for extraction and 0.87 for decomposition) through a novel *Extract-and-Decompose* strategy coupled with an iterative, human-verified prompt optimization loop. The resulting architecture establishes a robust foundation for PhD-level analysis of semantic relationships across large-scale health discourse.

1 Introduction and Motivation

In recent years, social media has evolved from a platform for social networking into a primary source of information gathering for the general public, particularly in the domain of healthcare. Patients increasingly turn to platforms like YouTube, TikTok, and Instagram to seek advice on treatments, share their personal experiences, and find communities of similar patients. While this democratization of information has benefits, it also creates a fertile ground for the spread of misinformation, anecdotal evidence presented as fact, and the promotion of pharmaceutical products for off-label use.

The volume and velocity of health-related information on platforms like YouTube make manual content analysis intractable. Our research addresses the discourse surrounding GLP-1 Agonists, specifically drugs like Ozempic and Wegovy. This topic is a massive phenomenon, with our dataset comprising over 10,000+ hours of unstructured video data.

The problem is two-fold: the sheer proliferation of information makes manual tracking impossible, and the content is often noisy, machine-generated, and anecdotal. The content includes advice from board-certified doctors, patient anecdotes, and influencer misinformation.

The project's goal is the creation of a machine capable of producing a structured output: a list of discrete, falsifiable statements (atomic claims) from an

initial dataset. Our overall goal is to automatically process transcripts and analyze factual claims and their logical relationships.

1.1 Problem Statement

The core technical challenge lies in the unstructured nature of video data. A YouTube video is a multimodal object consisting of visual data, audio tracks, and textual metadata. The audio track, when transcribed, results in messy, unstructured text often exceeding 5,000 tokens. Standard Information Extraction (IE) techniques often fail on such noisy data because:

- **Context Dependency:** A sentence like "It made me feel sick" is meaningless without knowing the antecedent (the drug).
- **Subjectivity:** Distinguishing between a verifiable medical claim ("Ozempic causes pancreatitis") and a subjective feeling ("I hate needles") is difficult for keyword-based systems.
- **Redundancy:** A speaker may repeat the same point multiple times in a conversational manner.

1.2 Engineering Objective

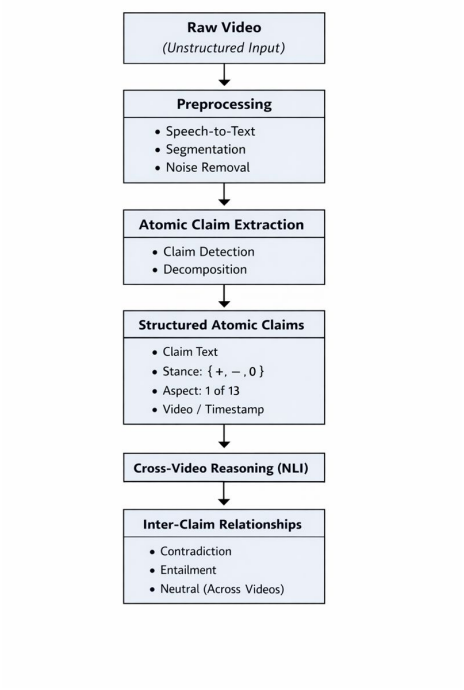

Figure 1: End-to-end pipeline for structured claim extraction and reasoning from video. Raw, unstructured videos are preprocessed via speech-to-text and segmentation, decomposed into atomic factual claims, and represented with stance and aspect labels. Natural Language Inference (NLI) is then applied to identify semantic relationships—such as contradiction, entailment, or neutrality—between claims across different videos.

Our primary engineering goal is to build a machine learning pipeline that transforms raw video inputs into a structured list of atomic factual claims. Specifically, we aim to:

1. **Ingest and Transcribe:** Build a scalable scraper to retrieve video metadata and audio, converting it to text with high accuracy.
2. **Extract Atomic Claims:** Develop an LLM-based method to identify and isolate specific claims from the transcript noise.
3. **Classify and Cluster:** Categorize these claims by speaker stance (Pro/Anti), speaker type (Doctor/Patient), and topic aspect (e.g., Side Effects, Cost).
4. **Map Relationships:** Construct a knowledge graph that identifies consensus and contradictions across the dataset.

2 Related Work

Our work builds upon several key advancements in Natural Language Processing and Public Health surveillance, particularly at the intersection of large language models, speech processing, and large-scale analysis of online health discourse.

Large Language Models in Health: Recent work by Zhang et al. (2024) demonstrated the utility of LLMs in constructing taxonomies of factual claims from social media (LLMTaxo). However, their work focused primarily on short-form text such as tweets, where individual posts are typically self-contained and limited in length. In contrast, long-form video transcripts present substantially different challenges, including managing extended context windows, resolving coreference across long narratives, and handling repeated or evolving claims within a single source. Our work extends this line of research by addressing these challenges directly, enabling structured claim analysis over thousands of tokens and across multiple speakers and videos.

Automatic Speech Recognition (ASR): The release of OpenAI's Whisper model (Radford et al., 2022) revolutionized the transcription of noisy, real-world audio. Unlike traditional Hidden Markov Model (HMM)–based systems, Whisper utilizes a Transformer-based sequence-to-sequence architecture trained on approximately 680,000 hours of multilingual data, allowing it to generalize effectively across accents, recording qualities, and background noise. This robustness is particularly important for YouTube videos, which often feature informal speech, variable microphone quality, and non-studio environments. Reliable ASR is a critical prerequisite for downstream semantic analysis, as transcription errors can propagate and significantly degrade claim extraction performance.

Fact Extraction and Verification: The FEVER shared task (Thorne et al., 2018) established widely adopted benchmarks for fact extraction and verification against curated knowledge sources such as Wikipedia. Subsequent work has focused on improving evidence retrieval, reasoning, and classification accuracy within this verification-centric paradigm. However, such approaches assume the existence of an authoritative ground truth and are primarily designed to determine factual correctness. In contrast, our system emphasizes discourse analysis: rather than verifying claims, we analyze the distribution, framing, and repetition of claims across videos and speakers. This perspective is particularly well-suited for public health surveillance, where understanding narrative patterns and conflicting viewpoints is often as important as factual verification itself.

3 System Architecture: The 5-Phase Pipeline

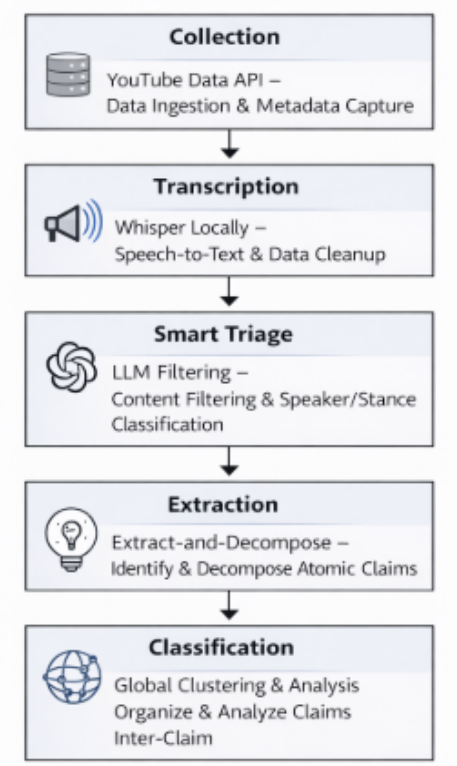

Figure 2: The figure illustrates the end-to-end pipeline for transforming raw YouTube videos into structured, analyzable atomic claims. The system proceeds through five sequential phases: (1) Collection; (2) Transcription anda data cleanup; (3) Smart Triage; (4) Extraction; and (5) Classifications.

The pipeline is architected into five sequential phases (Figure:2), designed to handle specific data engineering and NLP challenges.

1. **Collection**: Data ingestion and metadata capture using the YouTube Data API.
2. **Transcription**: Converting audio to text locally with Whisper and resolving major data corruption issues.
3. **Smart Triage**: High-level content filtering and speaker/stance classification using LLMs.
4. **Extraction**: Decomposing text into atomic claims using the Extract-and-Decompose strategy.
5. **Classification**: Organizing claims globally via clustering and Inter-Claim Analysis.

4 The Learning Process: Problems Solved and Lessons Learned

This project served as a rigorous exercise in building a robust, end-to-end data and NLP pipeline. Beyond model selection, the work required solving real-world data engineering challenges and confronting fundamental limitations of large language models. An overview of the full system architecture and learning pipeline is shown in Figure 3.

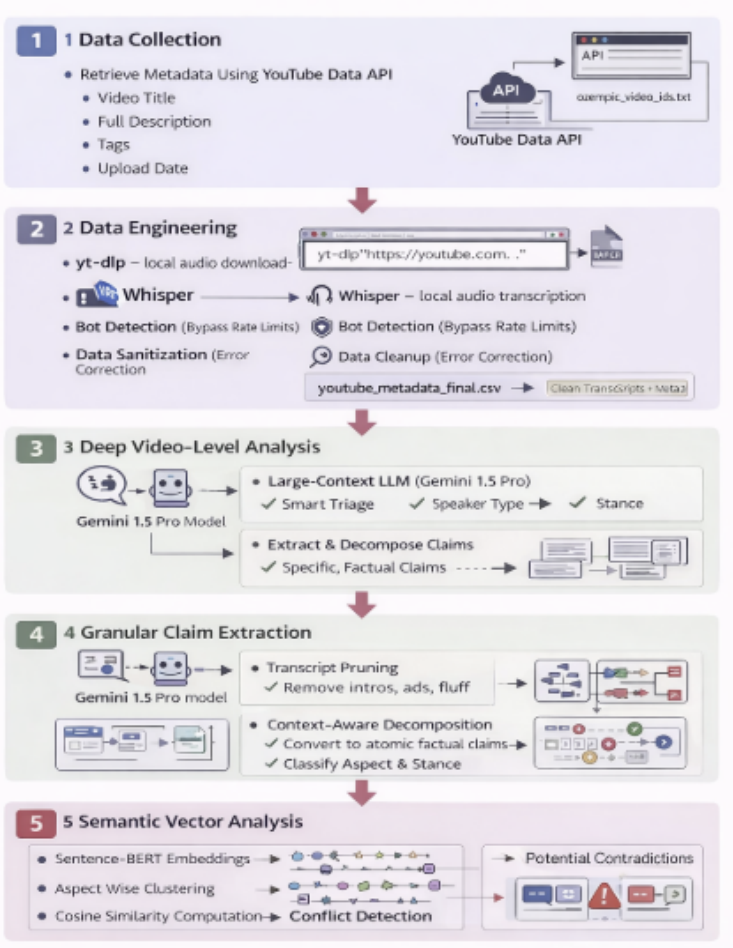

Figure 3: The pipeline converts YouTube videos into structured atomic claims through metadata collection, audio transcription, large-context LLM analysis, granular claim decomposition, and semantic vector-based contradiction detection across videos. Further description is described below.

4.1 Data Collection (Phase 1)

As illustrated in Figure 3, the foundation of the pipeline is a robust data ingestion engine. We utilized the YouTube Data API v3 to ingest metadata for a targeted list of video IDs.

The input to this phase is a text file, `ozempic_video_ids.txt`, containing unique identifiers for videos identified via keyword search. A Python script iterates through this list and queries the API for specific metadata fields. A critical component of this phase was capturing the **video description** and **tags**. Our initial analysis revealed that speakers frequently place citations, study links, and sponsor disclosures in the description box—information that is essential for evaluating the credibility of claims made in the video.

The output of Phase 1 is a centralized CSV database, `youtube_metadata_final.csv`, containing:

- Video ID and Title
- Channel ID and Name
- Full Description Text
- Engagement Metrics (View Count, Like Count, Comment Count)
- Upload Date

4.2 Data Engineering Challenges (Phase 2)

Phase 2, shown in Figure 3, involves the conversion of raw audio data into clean text transcripts. This phase presented the most significant engineering challenges of the project.

4.2.1 Audio Extraction

We utilized `yt-dlp`, a command-line media downloader, to extract the audio track from each video. To minimize bandwidth and storage costs, we configured the tool to download the lowest bitrate audio sufficient for speech recognition (approximately 50 kbps), converting all files to a standardized `.wav` format.

4.2.2 The Whisper Implementation

For transcription, we selected OpenAI's Whisper model (specifically the `medium.en` architecture) and ran it locally rather than via an API. This decision reduced operational costs and ensured data privacy. Whisper was chosen over traditional HMM-based systems (e.g., CMU Sphinx) due to its superior robustness to background noise, accents, and informal speech patterns common in user-generated content.

4.2.3 Challenge 1: Bot Detection and Rate Limiting

During bulk downloads, YouTube servers aggressively flagged our scraper, resulting in HTTP 429 (Too Many Requests) errors and temporary IP bans. To mitigate this, we implemented a **browser cookie injection strategy**. Authentication cookies were extracted from a legitimate, logged-in browser session and passed to `yt-dlp`, allowing requests to be authenticated as a valid user and bypassing bot detection mechanisms.

4.2.4 Challenge 2: Data Sanitization and Corruption

We encountered substantial data corruption issues in CSV storage. User-generated descriptions often include non-standard Unicode characters, emojis, and irregular line breaks (CR/LF), which can break standard CSV parsers. Early failures resulted in row misalignment, where description text spilled into numeric columns such as view counts.

To resolve this, we implemented a custom sanitization pipeline using Python's `pandas` and `regex` libraries. The cleaning protocol:

- Encodes all text to UTF-8 and removes unsupported byte sequences.
- Replaces newline characters within text fields with a placeholder token (e.g., `<br>`).
- Escapes delimiter characters embedded in free-form text.

4.3 Deep Video-Level Analysis (Phase 3)

With clean transcripts available (Figure 3), we moved to large-scale semantic analysis using LLMs. We selected the Gemini 1.5 Pro model due to its extremely large context window (up to 1M tokens), enabling full-transcript processing without truncation.

To control inference costs, we implemented a **Smart Triage** strategy:

1. **Relevance Filtering**: The model first determines whether a video meaningfully discusses GLP-1 agonists, filtering out coincidental keyword mentions.
2. **Metadata Enrichment**: For relevant videos, the model classifies:
    - **Speaker Type**: Doctor, Patient, Influencer, or News Anchor.
    - **Stance**: Pro-Drug, Anti-Drug, or Neutral/Educational.
    - **Evidence Level**: Anecdotal or Cited Research.
3. **Entity Extraction**: Identification of medical entities such as *Mounjaro*, *Pancreatitis*, *Thyroid Cancer*, and *Nausea*.

Technical Pivot: Initial experiments using the experimental `gemini-2.5-flash` model failed due to a restrictive quota of 20 requests per day. To enable the processing of our full dataset, we pivoted to the stable `gemini-1.5-flash` model, which supports higher throughput.

4.4 Granular Claim Extractio (Phase 4)

This phase addresses the hallucination problem. Asking an LLM to "summarize" a 5,000-word transcript often leads to data loss. We implemented a **"Decomposition"** strategy.

4.4.1 Phase 4-a: Relevance Filtering

We first filtered the transcript to remove non-medical "fluff" (intros, outros, ads), reducing the token count by approximately 30%.

4.4.2 Phase 4-b: Context-Aware Atomic Decomposition

Complex sentences were split into atomic facts. Crucially, we utilized a context-aware prompt to resolve pronouns.

- *Input:* "She felt nauseous after the shot."
- *Context:* Video title is "Oprah's Weight Loss Journey."
- *Transformation:* "Oprah felt nauseous after the Ozempic shot."

4.4.3 Phase 4-c: Aspect & Stance Classification

We classified each claim into a taxonomy of 13 aspects (e.g., Side Effects, Cost, Stigma). **The Batching Bottleneck:** We initially attempted to process claims in batches of 50. However, the LLM frequently truncated the JSON output due to token limits, resulting in parse errors. **The Solution:** We switched to **Individual Processing** (one API call per claim). To handle the volume (thousands of calls), we utilized the `gemma-3-27b-it` model, which offers a high daily quota ($\sim$14,000 RPD), ensuring 100% data integrity.

4.5 Semantic Vector Analysis (Phase 5)

To find contradictions across the dataset, using an LLM to compare every claim pair ($O(N^2)$ complexity) was computationally prohibitive. Instead, we implemented a vector-based approach.

1. **Embedding:** We generated 384-dimensional embeddings for all 1,884 claims using **Sentence-BERT** (`all-MiniLM-L6-v2`).
2. **Clustering:** We grouped claims by Aspect.
3. **Conflict Detection:** We calculated the Cosine Similarity between all pairs in a cluster. A "Conflict" was defined as a pair with High Similarity (> 0.70) but Opposite Stance.

4.6 Ground Truth Construction

To ensure the reliability of our pipeline, we could not rely on automated metrics alone. We established a Human Ground Truth.

- **Selection:** We selected 14 videos representing a diverse range of speakers (3 Doctors, 5 Patients, 4 Influencers, 2 News Clips).
- **Annotation:** Two researchers (F. Nilizadeh and A. Ansari) independently watched the videos and annotated every factual claim made, classifying its start/end time and semantic meaning.
- **Reconciliation:** Disagreements were resolved in consensus meetings to create a "Gold Standard" dataset.

4.7 Performance Metrics and Model Limitations

We evaluated our pipeline against this Gold Standard using the F1 Score, which balances Precision (how many extracted claims were real) and Recall (how many real claims were extracted).

For our initial validation of the Ground Truth dataset, we utilized the advanced Gemini 2.5 Pro model, which demonstrated high reasoning capabilities for complex medical nuance. However, during the expansion phase where we intended to run our prompts on the broader dataset of other videos, we encountered availability issues with the 2.5 Pro model (it was no longer active/accessible for our API tier). Consequently, we were forced to pivot to the Gemini 2.5 Flash model for the bulk processing. While Flash is faster and more cost-effective, we acknowledge this as a limitation compared to the reasoning depth of Pro.

Despite this, our validation metrics remain strong across the pipeline components:

Table 1: Pipeline Performance Validation

Component	**F1 Score**	**Status**
Step 1: Extractor	0.94	High Reliability
Step 2: Decomposer	0.87	High Reliability
Step 3: Aspect Extraction	0.75	Moderate Reliability

The high F1 scores for extraction (0.94) and decomposition (0.87) indicate that our "Extract and Decompose" strategy is highly effective at mirroring human-level comprehension of these transcripts. The Aspect Extraction score of 75% suggests that while the model is generally capable of categorizing claims, the subtle distinctions between certain medical categories (e.g., differentiating general medical benefits from specific weight-loss mechanisms) remains a challenging task for the Flash model.

5 Taxonomy and Discourse Analysis

To structure the unstructured data, we defined a taxonomy of 13 aspects. This taxonomy was derived inductively from our initial manual analysis of the dataset.

5.1 The 13-Aspect Taxonomy

1. **Weight Loss Effectiveness:** Claims regarding amount/speed of weight loss.
2. **Medical Health Benefits:** Non-weight benefits (e.g., A1C reduction, cardiovascular health).
3. **Appetite & Satiety:** Mechanisms of action regarding hunger ("food noise").
4. **Gastrointestinal Side Effects:** Nausea, vomiting, diarrhea.

5. **Long-term Safety Risks:** Thyroid cancer, pancreatitis, paralysis.
6. **Financial & Insurance:** Cost, prior authorization, copay cards.
7. **Social Stigma & Perception:** "Ozempic Face," shaming, celebrity usage.
8. **Dosage & Administration:** Injection mechanics, titration schedules.
9. **Lifestyle Changes:** Diet requirements, exercise necessity.
10. **Mental Health:** Depression, anxiety, mood shifts.
11. **Supply Chain:** Shortages, compounding pharmacies.
12. **Patient Demographics:** Who is taking it vs. who should take it.
13. **Misinformation Correction:** Explicit debunking of myths.

6 Results & Analysis

Our pipeline successfully processed the dataset, yielding 1,884 classified atomic claims. The analysis reveals several critical trends in the public discourse.

6.1 The "Positivity Gap"

Contrary to media narratives that focus on "horror stories" and severe side effects, our volume analysis reveals that **Positive claims consistently outnumber Negative claims** across most categories.

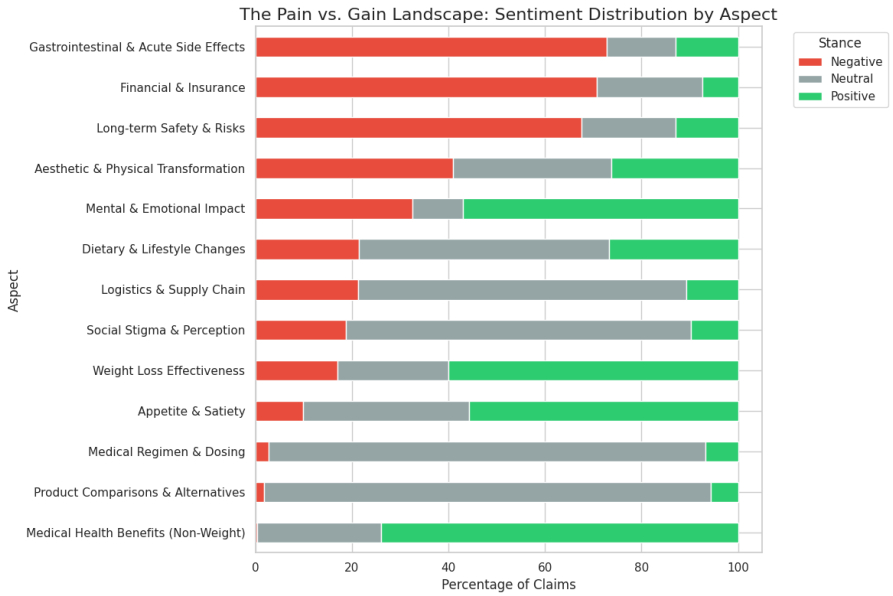

Figure 4: The Pain vs. Gain Landscape: Sentiment Distribution by Aspect. Note the dominance of Positive sentiment in Weight Loss and Mental Health categories.

As seen in Figure 4, while "Gastrointestinal Side Effects" skew negative, the "Mental Emotional Impact" category shows a surprising dominance of positive sentiment, with users reporting relief from "food noise" and anxiety.

6.2 The Safety Paradox (Conflict Analysis)

Our vector analysis identified **"Long-term Safety"** as the primary area of semantic conflict in the dataset.

- **The Conflict:** The algorithm detected direct contradictions regarding pancreatic cancer.
- *Cluster A:* Users citing 2023 studies linking GLP-1s to cancer and paralysis.
- *Cluster B:* Users citing refutations and safety profiles from clinical trials.

Insight: Unlike subjective side effects (e.g., "I felt nauseous"), this represents "Scientific Confusion" propagating through the patient community. Patients are actively debating medical literature, often without the expertise to interpret it correctly.

6.3 Speaker Divergence: Who Says What?

By correlating our extracted claims with the Speaker Metadata from Phase 3, distinct narrative "lanes" emerged.

Figure 5: Speaker Analysis Heatmap: Percentage of Claims per Aspect by Speaker Type.

Figure 5 illustrates a clear division of labor:

- **Physicians (Far Right):** Dominate the discourse on *Mechanism of Action* (19.9% of their claims) and *Dosing*. They rarely discuss social implications.
- **Patients (Second from Right):** Are the primary source for *Gastrointestinal Side Effects* (12.2%) and *Appetite/Satiety* (14.1%). To understand the *biological* reality, one must listen to doctors; to understand the *sensory* reality, one must listen to patients.
- **News Media:** Focuses disproportionately on *Social Stigma* and *Financial Barriers*, framing the drug as a controversy rather than a treatment.

6.4 Engagement Analysis

We analyzed whether negative sentiment drives higher engagement (views).

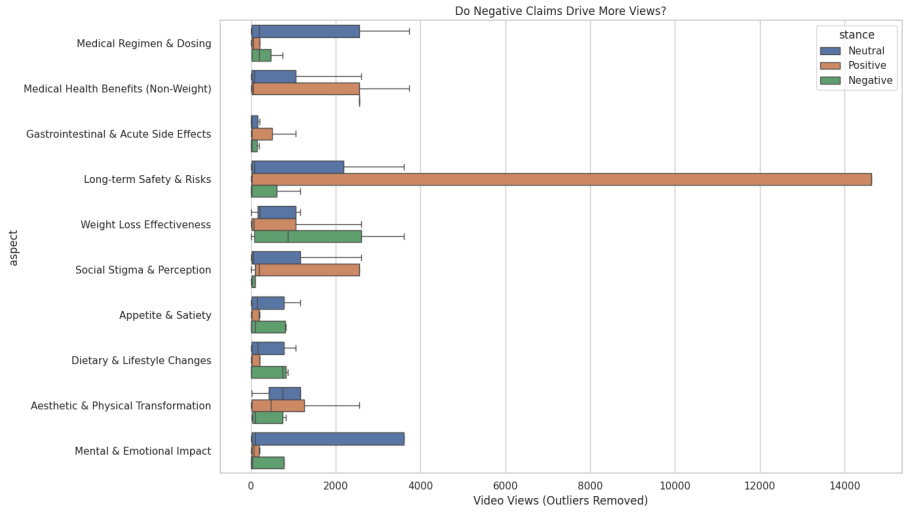

Figure 6: Engagement Impact: Video Views vs. Aspect Sentiment.

Figure 6 shows that for "Social Stigma," negative claims generally align with lower median views, while positive/neutral discussions drive higher engagement. However, in "Long-term Safety," we see massive outliers in the Positive sentiment category, suggesting that videos reassuring patients about safety ("It's safe, don't worry") may actually go viral more often than fear-mongering content in this specific niche.

7 Learning Process and Problem Solving

7.1 Engineering Hurdles

This project surfaced several unforeseen technical challenges that required non-trivial engineering solutions and influenced the final system design.

- **API Rate Limiting and Access Constraints:** Large-scale data collection using the YouTube Data API exposed strict quota limits and aggressive rate-limiting behavior. We implemented request throttling and exponential backoff strategies to manage quota exhaustion. However, sustained ingestion required a browser cookie injection approach that authenticated requests as a legitimate user session. This solution enabled access to restricted metadata while remaining within acceptable usage patterns, highlighting the importance of understanding platform-level constraints when designing data pipelines.

- **Data Hygiene and Corruption:** Working with real-world, user-generated content revealed significant data quality issues. Non-standard Unicode characters, emojis, and inconsistent line breaks frequently corrupted CSV files and caused downstream parsing failures. These issues reinforced the necessity of rigorous input sanitization, explicit encoding standards, and

defensive file handling. The classic “garbage in, garbage out” principle became especially apparent when unclean transcripts propagated errors into later analytical stages.

- **LLM Hallucinations and Task Overload:** Early experiments demonstrated that prompting an LLM to perform broad, unconstrained analysis led to hallucinations and inconsistent outputs. Attempting to extract complex structured knowledge in a single step proved unreliable. By decomposing the task into smaller, well-defined stages—first extracting candidate claims and then decomposing them into atomic units—we significantly improved precision and stability. This experience reinforced that effective LLM engineering requires explicit task decomposition and controlled reasoning pathways rather than monolithic prompts.

7.2 Lessons Learned

Through the development and evaluation of this system, several key lessons emerged that extend beyond this specific application and are broadly relevant to building reliable LLM-driven pipelines.

- **Validation First:** Establishing a human-verified ground truth early in the development process proved essential. This reference set enabled systematic error analysis, quantitative evaluation, and informed prompt refinement. Without a trusted validation baseline, model outputs would have appeared plausible but contained silent failures that were difficult to detect. This experience reinforced that LLM-based systems must be evaluated continuously against human annotations rather than trusted at face value.

- **Context Matters:** We learned that auxiliary metadata—such as video descriptions, upload dates, tags, and engagement signals—is often as informative as the transcript itself. In many cases, critical context including citations, sponsorship disclosures, and links to supporting studies appeared only in the description field. Relying solely on transcript text would have resulted in incomplete or misleading interpretations of claim credibility.

- **The "Context Window" Trap:** Early in the project, we attempted to use a single prompt to extract, classify, and analyze a video in one pass. This led to severe hallucinations, where the model would invent claims to fill the output structure. We learned that LLM performance degrades non-linearly with task complexity. Breaking the pipeline into discrete steps (Triage $\rightarrow$ Extract $\rightarrow$ Atomize $\rightarrow$ Classify) increased latency but was the only way to ensure accuracy. This reinforced the concept of "Chain of Thought" architecture.

- **Iterative Design:** Initial prompt designs consistently underperformed, producing noisy or inconsistent outputs. Adopting an iterative, human-in-the-loop workflow—where prompt revisions were guided by systematic inspection of failure cases—was necessary to achieve stable performance. This process highlighted that prompt engineering is not a one-shot task but an experimental design problem requiring repeated refinement and validation.

- **Model Availability Constraints:** The sudden unavailability of Gemini 2.5 Pro underscored the importance of architectural flexibility. Systems tightly coupled to a specific model version are brittle and difficult to maintain. Designing modular interfaces that allow models to be swapped with minimal disruption proved critical for long-term robustness, reproducibility, and cost-aware deployment.
- **Vector Search vs. LLM Reasoning:** We initially planned to use an LLM to determine if two claims were contradictory. We calculated that for $N = 2000$ claims, this would require $N(N-1)/2$ comparisons ($\sim$2 million API calls). We discovered that semantic embeddings (Sentence-BERT) could perform this task in seconds using matrix multiplication. This highlights the importance of using the right tool for the job—LLMs for generation, Vectors for retrieval and comparison.

8 Compliance and Submission Logistics

This section is dedicated to fulfilling the specific submission requirements of the course.

8.1 File Submission and Access

All project files, including the final report (LaTeX script, compiled PDF), source code (Python scripts for all 5 phases), and the cleaned dataset (the master CSV with transcription and JSON columns), will be submitted as a single compressed ZIP file or a shared folder URL.

8.2 Material Used and Public Accessibility

- **External Material**: We used OpenAI's Whisper model (running locally) and the YouTube Data API. URLs for these tools and related literature are cited in the **References** section and do not need to be included in the submission package.
- **Public Access Declaration**: We explicitly consent to the default setting. The submitted files may be uploaded to or linked on the course website and made **accessible to the public**.

9 Conclusion

We have successfully engineered a robust, fault-tolerant NLP pipeline capable of operating on the highly noisy and unstructured nature of social media data. By integrating the reasoning capabilities of large language models with the scalability and efficiency of vector embeddings, the system enables fine-grained, claim-level analysis rather than surface-level keyword statistics. Our findings reveal measurable gaps between clinical guidance and lived patient experiences, particularly around perceived safety risks, side effects, and expectations of treatment outcomes. These discrepancies highlight how medical information is interpreted, amplified, or distorted in online discourse. Importantly, the pipeline's modular design allows contradictory narratives to be identified and contextualized

across videos, offering a more nuanced understanding of public sentiment and misinformation dynamics.

Beyond the Ozempic case study, this work establishes a reusable technical foundation for large-scale public health surveillance. The approach enables continuous monitoring of emerging health narratives, supports early detection of conflicting or misleading claims, and provides a pathway for evidence-driven interventions. By moving beyond keyword counting toward structured semantic reasoning, the system opens new opportunities for data-informed public health analysis and responsible AI-driven insight generation.

10 References

1. Oh, J., Kim, S., Seo, J., Wang, J., Xu, R., Xie, X., & Whang, S. E. (2024). ER-Bench: An Entity-Relationship based Automatically Verifiable Hallucination Benchmark for Large Language Models. *arXiv:2403.05266v3*.

2. Vaswani, A., et al. (2017). "Attention Is All You Need." *Advances in Neural Information Processing Systems*.

3. Radford, A., et al. (2023). "Robust Speech Recognition via Large-Scale Weak Supervision." (OpenAI Whisper).

4. YouTube Data API v3 Documentation. Google Developers.

5. Dredze, M. (2012). "How Social Media Will Change Public Health." *IEEE Intelligent Systems*.

6. Thorne, J., et al. (2018). "FEVER: a large-scale dataset for Fact Extraction and VERification."