POLITECNICO DI TORINO

Master's Degree Programme
in Mathematical Engineering

Master's Thesis

A Data-Driven Investigation of Potentially Harmful Diet-Related Content on YouTube: Detection and User Engagement Analysis

Internal supervisor
Prof. Tatiana Tommasi

Candidate
Maddalena Ghiotti

External supervisors
Prof. Daniela Paolotti
Dr. Yelena Aleksandrovna Mejova

Academic Year 2025–2026

Abstract

In recent years, social media have become popular sources of information, including in the context of nutrition and weight loss. However, the spread of low-quality content that emphasizes physical appearance poses a risk to viewers' self-esteem, body image, and body satisfaction, and in the most severe cases, it may contribute to the development of disordered eating behaviors (DEBs) or, potentially, eating disorders (EDs).
This study analyzes 3129 YouTube videos about diet and weight loss, considering titles, descriptions, transcripts, and metadata. The analysis aims to explore the relationship between these metadata, video topics, and the level of risk associated with each video, to facilitate the detection of harmful material and inform moderation practices.

We quantify the risk using 23 principles derived from three questionnaires. Information quality is assessed through the Principles for Health-related Information on Social Media (PRHISM) and the Health on the Net Foundation Code of Conduct (HONcode), while body-related content is measured using the body-related variables proposed by Munro et al. Each principle is scored on a 0–4 scale via prompts to the GPT API, and the resulting scores are validated on a human-labeled sample, on which the extent of AI use and the presence of brand or branded product mentions are also manually evaluated. Concurrently, non-negative matrix factorization (NMF) is applied to identify 13 thematic dimensions and map each video to one or more of them, based on dimension-dependent thresholds. Beyond scores and topics, non-semantic variables are extracted, including the channel age and the number of mentions in the video description.
Correlation analyses, statistical tests, and linear regression models were adopted to examine the relationships between quality and body scores and engagement metrics, as well as between non-semantic variables and video topics, on the one hand, and quality or body scores, on the other. Predictive performance was further assessed using simple neural network models trained to estimate these scores from the aforementioned variables.

The two variables introduced above are significantly positively associated with video quality, while quality decreases as the proportion of uppercase letters in titles and descriptions increases. Surprisingly, the topics *Personal storytelling* and *Mindset & Motivation* are also linked to higher-quality content, whereas negative associations emerge for *Supplement review* and *Mounjaro recipe*. *Mitolyn* and *Mounjaro* are among the most frequently cited words, revealing a suspicious pattern: they are predominantly mentioned in videos categorized as *Music* (by the creator) and published by South American channels, often featuring the same individuals across different accounts. Videos in this category and region are more likely to be of lower quality. Overall, non-semantic variables and video topics provide a good explanation of the quality score ($R^2_{adj} = 0.58$). Their explanatory power is lower for the body score ($R^2_{adj} = 0.39$), although video topics still contribute substantially. In this case, *Personal storytelling* and *Mindset & Motivation* are associated with higher body scores, as are videos self-declared under the *Health* category. No significant patterns emerge in the relationship between quality and body scores and engagement metrics.

The results show that topics and non-semantic variables aid in identifying potentially harmful diet-related content, providing insights for automated detection and moderation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In recent years, the consumption of news and information through social media (SM) has increased significantly, profoundly changing the ways in which users access content. Public attention toward who communicates the news is also shifting: on TikTok, Instagram, and Snapchat, users mainly follow celebrities, influencers, and creators, whereas on Facebook and X/Twitter, traditional news outlets remain central in public discourse (Digital News Report [2023]). At the same time, visual and video-based platforms, such as YouTube and TikTok, are playing an increasingly important role, while more traditional SMs like Facebook are experiencing a decline in informational use (Digital News Report [2023]). In particular, YouTube is perceived as a platform useful for learning and personal development, whereas TikTok and Instagram are more often associated with entertainment or short-form informational content (Horning [2024]).

In this context, YouTube stands out as one of the main platforms for the dissemination of informational content. User-generated content (UGC) on YouTube facilitates the circulation of news and increases audience engagement, resulting in higher news consumption compared to traditional sources (Chunqiong et al. [2025]). YouTube's relevance is even more evident when looking at younger users: according to the Pew Research Center, approximately nine out of ten U.S. teenagers use the platform, making it the most widely used among those considered (Anderson et al. [2023]). Not only is it the most widely used, but 71% of teenagers access it daily, and 16% report using it almost constantly. TikTok, Snapchat, and Instagram remain popular but at lower rates (Anderson et al. [2023]).

Although many users do not actively seek news on YouTube or TikTok, informational content often appears in feeds and recommended sections as an extension of other content they are viewing (Horning [2024]). However, algorithmic influence can negatively affect users' trust (Chunqiong et al. [2025]), which is generally declining: only four out of ten people report trusting most news most of the time (Digital News Report [2023]). Nevertheless, users still show a slight preference for content selected algorithmically based on their previous interactions over content chosen by journalists or editors (30% vs. 27%) (Digital News Report [2023]), and trust in UGC is higher compared to traditional sources (Chunqiong et al. [2025]).

The increasing spread of informational content is particularly evident in the context

of nutrition and weight loss. The NutriNet-Santé study showed that 85.1% of interviewed French adults used the Internet to search for health and/or nutrition-related information, while 13.6% reported reading or posting messages on online health and/or nutrition forums (Fassier et al. [2016]). In Saudi Arabia, 89.2% of surveyed adults declared using social media to obtain weight-loss information (Alzaben et al. [2022]). Similarly, a survey conducted in South Africa (Kreft et al. [2023]) revealed that 69.7% of participants used social media to access nutrition-related information, and among them, 96% identified YouTube as the most frequently used platform. This further confirms the relevance of YouTube in this context, particularly when nutrition and weight loss are concerned.

Nutrition- and body-related content is often presented as a fundamental basis for success and well-being, frequently encouraging strict dietary routines or restrictive eating behaviors. Overall, the belief that thinness, stereotypical fitness, and physical appearance are more important than health is commonly referred to as *diet culture*. Within this framework, food choices are often motivated by non-scientifically validated information, presented as optimal strategies for achieving fitness or health. Such content is frequently associated with hashtags such as #WhatIEatInADay and #CleanEating. Other common terms include *fitspiration* and *thinspiration*, which predominantly focus on physical appearance and promote extremely thin or extremely fit bodies as the sole desirable standards. Females represent the primary contributors to the creation of nutrition-related content on social media, and in four out of ten most-viewed YouTube videos on this topic, they display their bodies at least once. Moreover, most of this content promotes a weight-normative view of health, while weight-inclusive perspectives appear only marginally (Minadeo and Pope [2022], Basch et al. [2017]).

Regarding information-seeking behavior, only a minority of users (17%) actively search for nutrition-related content on social media, while the majority (54%) passively encounter such content through their feeds, especially women (Mayoh and Jones [2021]). As a consequence, both youths and adults are frequently exposed to videos and posts created by non-certified content creators and disseminated outside accredited information sources. Institutional or authoritative sources represent only a small fraction of the overall available content, and viewers rarely feel the need to further verify the validity of the information encountered online. In fact, only 16.0% of NutriNet-Santé study participants reported discussing the information found online with a healthcare professional. This issue is further amplified among individuals with lower educational levels or limited digital skills (Fassier et al. [2016]), who are often characterized by lower income and reduced access to healthcare support. Even when users wish to assess credibility, determining the accuracy of nutrition-related information on social media remains a challenging task (Kreft et al. [2023]), rendering this audience particularly vulnerable.

There is currently no universally agreed-upon evidence demonstrating a direct causal relationship between social media use and body image disturbance, and only a limited number of studies identify a small positive association between the two (Saiphoo and Vahedi [2019], Holland and Tiggemann [2016]). The amount of time spent on social media, as well as the number of platforms used, does not appear to be significantly associated with body image disturbance or disordered eating behaviors. However, the type of content consumed plays a crucial role, with appearance-focused content showing a stronger effect

(Saiphoo and Vahedi [2019], Sanzari et al. [2023]).

Body image has been defined by Thompson et al. [1999] as "the internal representation of one's own outer appearance — one's unique perception of the body." Body image disturbance is a collective term referring to dysfunctions affecting emotional, cognitive, behavioral, or perceptual dimensions of this representation. Within this framework, body dissatisfaction represents a specific and fundamental component, described as the "most important global measure of distress," and refers to dissatisfaction with specific body aspects. Disordered eating behaviors (DEBs) are instead defined as "troublesome eating behaviors, such as purgative practices, binge eating, food restriction, and other inadequate methods to lose or control weight" (Pereira and Alvarenga [2007]). Eating disorders (EDs) should not be confused with DEBs. According to the National Institute of Mental Health (NIMH), EDs are "serious illnesses marked by severe disturbances in eating behaviors," whereas DEBs are not classified as illnesses and occur with lower frequency or severity than required for an ED diagnosis.

Exposure to fitspiration content has been shown to increase body dissatisfaction, physical appearance comparisons, and negative mood, while decreasing state appearance self-esteem (Jeronimo and Carraca [2022], Tiggemann and Zaccardo [2015], Pearl and Puhl [2016]). Furthermore, sentiment analysis conducted by Tiggemann et al. [2018] reported even higher negativity levels in content related to *thinspiration*. Exposure to weight-loss content or to body positivity/neutrality content has also been associated with lower body appreciation, greater fear of negative appearance evaluation, and more frequent binge eating and laxative use, as demonstrated by Sanzari et al. [2023] using the Body Appreciation Scale (BAS) and the Fear of Negative Appearance Evaluation Scale (FNAES). The same study showed that participants interviewed in 2022 reported higher frequencies of vomiting and laxative use compared to those interviewed in 2015, a trend paralleling the increase in YouTube usage. The frequency of binge eating and laxative use was also higher among women.

Overall, several studies (Minadeo and Pope [2022], Fassier et al. [2016], Pearl and Puhl [2016]) conclude that exposure to nutrition-related online content can contribute to the development of disordered eating behaviors, particularly among individuals who engage in dieting motivated by appearance-related goals.

Age-related differences in body dissatisfaction are evident, with younger individuals being more vulnerable. In these age groups, the association between social media use and body image disturbance appears to be stronger (Davey et al. [2024], Jeronimo and Carraca [2022], Saiphoo and Vahedi [2019]). Differences by age are particularly pronounced in the context of eating disorders. Study Davey et al. [2024] demonstrated that differences in Eating Disorder Examination Questionnaire (EDE-Q) scores following exposure to #CleanEating and #WhatIEatInADay content, compared to a control condition (#Nature), were more marked among younger participants aged 18–21.

According to several studies (Sharma and Vidal [2023], Dahlgren et al. [2024], Nawaz et al. [2024]), eating disorders are also associated with social media use, particularly image- and video-based platforms and nutrition-related, appearance-centered content. This association appears stronger among females, who tend to internalize the thin ideal, while males are more often associated with the muscular ideal. Nevertheless, research on this

topic remains limited, and consensus has not yet been reached. As reported in the literature review by Dane and Bhatia [2023], "Eight studies investigated the impact of the fitspiration trend on body image dissatisfaction and eating disorder pathology with mixed results: 50% supported the relationship, 25% partly supported it, and 25% refuted it."

Due to the described context, it has become essential to implement preventive measures. These measures should be undertaken both by SM platforms and by health authorities committed, among other objectives, to scientific dissemination. To design appropriate interventions, it is first necessary to understand the current informational environment and to identify which types of content pose the greatest risk for young users.

The aim of this study is therefore to identify the characteristics of YouTube channels and videos related to dieting and weight loss that may steer young people toward eating disorders or unhealthy attitudes toward food or body. Additionally, we aim to assess the reach and popularity of these high-risk contents, with the goal of providing quantitative evidence.

This investigation began with two central questions:

- Which types of content pose the greatest level of risk?
- What is the level of engagement associated with the highest-risk videos?

Since there is no objective or universally accepted measure of "risk" in this context, and because risk is influenced by a wide range of factors that would be difficult to fully capture within the scope of this thesis, the study focuses on two elements identified in the reviewed literature as key contributors: low informational content quality (including AI-generated or AI-modified content) and high levels of body-related content.

Other important aspects, such as the scientific accuracy of the information presented, are not included in the analysis due to their limited measurability and the constraints of the project, including time and resources. As a result, we reframed the original goal of quantifying risk as a whole into the study of two specific components that can be considered part of the broader construct of risk.

We also included a third, more minor but still measurable factor: the presence of product or brand mentions, which may play a role in shaping viewer vulnerability or posting party's incentives, moving towards the interests of a company, rather than of the audience.

We therefore refined the research questions as follows:

Q1. What are the main topics covered by dieting and weight-loss videos on YouTube?

Q2. Which topics are associated with higher levels of body-related content or lower informational quality? Additionally, to what extent are non-semantic characteristics, including channel features, product or brand mentions, and the presence of AI-generated content, associated with differences in body-related content or informational quality?

Q3. How do the different content characteristics (body-related, informational quality, video topics, and non-semantic metadata) relate to engagement levels?

Q4. To what extent can the level of body-related content or informational quality be accurately predicted using topic features and non-semantic characteristics through standard and low-cost models?

To address these research questions, we employed multiple methodological approaches, including topic modeling techniques, manual and LLM-assisted video labeling, as well as statistical models and hypothesis testing procedures. This data-driven framework, leveraging algorithmic methods and large language models, enabled an in-depth analysis of a substantially larger sample compared to previous studies, thereby enhancing the robustness and generalizability of the findings.

Furthermore, focusing specifically on diet and weight loss discourse on YouTube contributes to filling a gap in the existing literature, mainly focused on other SMs and platforms, and provides novel tools for risk assessment within this domain.

In this context, statistical methodology played a central role, ensuring objectivity and validity of the results and making the study particularly aligned with the analytical and quantitative foundations of the Mathematical Engineering program.

The remainder of this thesis is structured as follows.

Chapter 2 reviews the relevant literature, covering topic modeling for unstructured transcripts, standardized frameworks for evaluating digital health information quality, and the use of Large Language Models (LLMs) in automated content assessment.

Chapter 3 details the study's methodology. It outlines the YouTube data collection and pre-processing pipeline, the application of Non-Negative Matrix Factorization (NMF) for topic extraction, and the adaptation of established questionnaires to measure information quality and body-related content. Furthermore, it describes the LLM-based automated annotation process and the analytical framework used for statistical and predictive modeling.

Chapter 4 presents the empirical results. After exploring the dataset's descriptive statistics and the identified thematic landscape, it validates the LLM-scoring approach against human annotations. The chapter then addresses the core research questions by analyzing the determinants of video quality and body-related content, their relationship with user engagement, and the performance of predictive neural network models.

Finally, chapter 5 concludes the work by summarizing the key findings and discussing their implications for automated moderation practices. It also critically addresses the study's limitations and outlines promising directions for future research.

Chapter 2

State of the art

The digital landscape has transformed how individuals access information about health, nutrition, and weight loss. Social media platforms such as YouTube are now major sources of diet-related content, raising concerns about misinformation, unrealistic body ideals, and the potential promotion of disordered eating behaviors and eating disorders.

Previous literature has extensively documented the prevalence of such risks. Exploratory content analyses, such as those conducted by Basch et al. [2017] and Tang et al. [2022], have highlighted how weight-loss videos on YouTube frequently promote rapid, unsustainable dietary changes and generate massive user engagement. To systematically assess the informational value of these trends, recent studies have employed standardized frameworks. For instance, Denniss et al. [2024] utilized the Principles for Health-related Information on Social Media (PRHISM) to demonstrate the generally poor quality of nutrition advice shared by influential Instagram accounts. Similarly, Zeng et al. [2025] evaluated the accuracy and engagement of dietary content on TikTok, revealing a concerning proliferation of weight-normative messaging lacking expert oversight. Concurrently, the psychological risks associated with this media have been investigated by Munro et al. [2024], who developed specific coding schemas to capture the prevalence of potentially harmful appearance-focused behaviors, such as *body checking* and extreme caloric restriction, in short-form videos.

Despite these valuable contributions, a significant empirical gap remains in the literature. The vast majority of recent research focusing on nutrition, diet culture, and body image has predominantly targeted highly visual, short-form platforms such as Instagram, TikTok, and Twitter. In contrast, YouTube, despite being the largest video-sharing platform globally and a primary source for health-related searches, remains surprisingly under-researched in the specific context of diet and weight loss. Furthermore, most existing studies evaluating information quality and body-related risks rely heavily on manual annotation procedures (Denniss et al. [2024], Munro et al. [2024], Zeng et al. [2025]). While accurate, manual coding is extremely labor-intensive and inherently restricts the scope of the research to small sample sizes (typically ranging from 100 to 500 videos or posts). Consequently, it is difficult to achieve a comprehensive, large-scale mapping of the thematic landscape and to robustly model how multiple dimensions of risk, namely, low informational quality and high physical appearance focus, interact with user engagement

across thousands of multimedia items. Furthermore, while the impact of semantic and non-semantic video features (e.g., duration, channels metadata) on audience attention has been modeled in general domains (Dai and Wang [2023]), a predictive synthesis linking these variables to specific dietary risks is lacking.

This thesis aims to fill these gaps by proposing a highly scalable, data-driven framework to detect and analyze potentially harmful diet-related content on YouTube. Building upon recent methodological advancements demonstrating the efficacy of Large Language Models (LLMs) as medical content evaluators (Khalil et al. [2025]), this study replaces the bottleneck of manual coding with a rigorous Zero-Shot prompting pipeline powered by GPT-4.1. By adapting established questionnaires (PRHISM, HONcode, and Munro et al.'s variables) into structured LLM prompts, this research automatically scores both the informational quality and the level of body-related content across a massive dataset of over 3,000 YouTube video transcripts.

By integrating this automated annotation with Non-Negative Matrix Factorization (NMF) topic modeling and the extraction of non-semantic metadata, this work provides a comprehensive overview of the diet and weight-loss ecosystem on YouTube. Ultimately, it uncovers the underlying relationships between thematic choices, video quality, and user engagement, offering novel predictive models that can inform and enhance automated moderation practices for public health safety.

The following sections review the main methodological approaches that support the analytical framework adopted in this thesis. Given the multidisciplinary nature of the study, the literature spans several areas.

First, section 2.1 on topic modeling discusses techniques used to extract thematic structures from textual data, with particular attention to approaches suitable for noisy sources such as video transcripts. Both traditional models, such as Latent Dirichlet Allocation and Non-Negative Matrix Factorization, and more recent neural approaches based on contextual embeddings are considered.

The following section (section 2.2) examines frameworks developed to assess the quality and reliability of online health information, as well as recent efforts to evaluate such content specifically within social media environments. Particular attention is given to the PRHISM framework, which is designed to assess health-related information shared on social platforms.

Subsequent sections address methods used to identify appearance-focused or body-related content in social media videos (2.3) and review established practices for manual annotation and the assessment of inter-rater agreement (2.4). Finally, recent literature on the use of Large Language Models as automated evaluators is discussed, highlighting their growing role in large-scale content annotation tasks (2.5).

2.1 Topic modeling for transcript analysis

Understanding the thematic structure of video transcripts is an important step in the analysis of large collections of online video content. Identifying the main topics discussed in these transcripts allows researchers to characterize video themes in a systematic and scalable way, enabling subsequent analyses that relate content categories to other video

features. For this reason, topic modeling has emerged as a transformative computational technique for discovering latent thematic structures within large, unstructured textual corpora (Grootendorst [2022]). When dealing with conversational data, such as audio and video transcripts, researchers face unique challenges: these texts are often lengthy, highly unstructured, polythematic, and susceptible to speech-to-text transcription errors (Cheng et al. [2022], Thies et al. [2021]).

Latent Dirichlet Allocation and Non-Negative Matrix Factorization

Historically, probabilistic models such as Latent Dirichlet Allocation (LDA) have dominated the field. LDA treats documents as mixtures of topics and topics as distributions over words, operating strictly on a bag-of-words (BoW) assumption. This traditional approach has been widely successfully applied in social media analysis; for instance, Saura et al. [2020] combined LDA with sentiment analysis to explore user-generated content on Twitter regarding healthy diets and food categories. However, LDA is not well suited for our corpus because the documents include noisy speech transcriptions belonging to a single macro-topic, conditions under which frequent conversational words may dominate the bag-of-words representation and lead to weakly distinguishable topics, or the creation of a generic "background" topic dominated by common words. As an alternative, Non-Negative Matrix Factorization (NMF) has proven to be highly effective, thanks to its application on the TF–IDF document–term matrix, which helps reduce the influence of very frequent terms shared across documents. NMF decomposes the matrix into two lower-rank non-negative matrices (a dictionary and a coding matrix), producing a naturally sparse and highly interpretable parts-based representation (Cheng et al. [2022]). Unlike some hard-clustering methods, NMF offers the flexibility to assign multiple broad topics to a single document, capturing the polythematic nature of complex texts.

The integration of contextualized neural models

While NMF excels at identifying interpretable, broad themes, traditional implementations may occasionally miss fine-grained semantic nuances because they do not incorporate contextual word embeddings (Cheng et al. [2022]). Consequently, the field has seen a surge in neural topic models, most notably BERTopic (Grootendorst [2022]). BERTopic leverages pre-trained transformer models (such as Sentence-BERT) to generate dense contextual embeddings, reduces their dimensionality using UMAP, and clusters them via HDBSCAN before extracting topic representations through a class-based TF-IDF (c-TF-IDF) procedure (Grootendorst [2022]).

The superior contextual understanding of neural models has led to their adoption in various spoken-language domains. For instance, Arfaoui et al. [2025] successfully applied BERTopic to focus group transcripts, demonstrating that contextualized embeddings capture semantic nuances in multi-party dialogues significantly better than traditional LDA. Similarly, Lalk et al. [2024] utilized BERTopic on psychotherapy session transcripts to predict clinical metrics like therapeutic alliance and symptom severity. In the realm of digital health, Zhang et al. [2024] combined state-of-the-art Automatic Speech Recognition (OpenAI's Whisper) with BERTopic to automatically identify depression-related

topics from smartphone-collected free-response speech recordings. Furthermore, Stöckl, A. utilized a pipeline of Whisper, GPT-3 for keyword summarization, and BERTopic to perform dynamic topic modeling on Austrian TV commentaries, effectively mapping the temporal evolution of dominant political and social themes.

Hybrid and Graph-Based innovations for video transcripts

Despite the advanced semantic capabilities of neural models like BERTopic, they present notable drawbacks when applied directly to full video transcripts: they typically restrict a document (or sentence) to a single topic and often produce an unmanageably fragmented number of subtopics (Cheng et al. [2022]).

To harness the complementary strengths of both paradigms, Cheng et al. [2022] proposed a Multi-Scale Hybridized Topic Modeling (MSHTM) approach. In this pipeline, NMF is first applied at the macro-level to accurately partition full interview transcripts into broad, overlapping categories. Subsequently, BERTopic is deployed exclusively on the sentence level within those NMF-defined clusters to uncover hidden, highly specific subtopics. This hybridization significantly lowers the computational and memory costs compared to running BERTopic on the entire corpus, while preserving NMF's vital ability to assign multiple macro-topics to complex spoken responses.

Other researchers have explored alternative unsupervised methodologies to handle the specific noise inherent in YouTube auto-generated transcripts. For example, Thies et al. [2021] introduced GraphTMT, an approach that models the vocabulary as a graph where edges are weighted by the cosine similarity of word embeddings. By extracting k-component subgraphs, GraphTMT filters out the noise caused by speech-to-text errors and extracts coherent topics without requiring the user to predefine the exact number of clusters, a distinct advantage over traditional distance-based algorithms when exploring unknown video datasets (Thies et al. [2021]).

Collectively, these advancements demonstrate that while neural embeddings provide deep semantic context, matrix factorization frameworks like NMF remain structurally essential for accurately mapping the multi-thematic reality of long-form video transcripts.

2.2 Evaluating information quality: frameworks and assessment tools

In the context of online videos discussing diet and weight loss, although not limited exclusively to this domain, assessing the quality of the information provided is essential to identify potentially harmful content. Incomplete, or misleading information may expose viewers to health risks, making informational quality an important component in the evaluation of the overall risk associated with such material.

In response to these concerns, the scientific community has developed standardized methods to evaluate the reliability and overall quality of digital health information. Early efforts in this field were primarily designed for static websites and written educational materials, focusing on aspects such as the credibility of the publisher and the scientific soundness of the information provided.

One of the earliest and most widely adopted tools is the DISCERN instrument, originally formulated to help patients and professionals assess the quality of written consumer health information regarding treatment choices (Charnock [1998]). DISCERN utilizes a 16-item questionnaire divided into three sections: reliability of the publication, quality of the specific treatment information, and an overall quality rating. Due to its robustness, it has been extensively applied across various domains of digital health. For instance, researchers have employed DISCERN to evaluate the reliability of websites describing the Mediterranean diet (Hirasawa et al. [2012]), as well as related content on YouTube (Benajiba et al. [2023]). The instrument has also been applied to assess online resources providing dietary recommendations for kidney stone formers (Traver et al. [2009]) and to evaluate the educational value of YouTube videos providing nutritional information for patients after bariatric surgery (Batar et al. [2020]).

Alongside DISCERN, the EQIP (Ensuring Quality Information for Patients) tool was developed by Moult et al. [2004] to measure the presentation quality of written healthcare data. Unlike other instruments, EQIP comprises 20 items that not only evaluate the understandability and layout of the content but also prescribe specific actions to be taken (e.g., whether to retain, revise, or discard a leaflet) based on the final score. In comparative web evaluations, EQIP and the first section of DISCERN have often been used in tandem to gauge both presentation quality and underlying reliability (Gkouskou et al. [2011]).

To further standardize medical content on the web, ethical codes and benchmarks were established. The Health On the Net Foundation Code of Conduct (HONcode) and the Journal of the American Medical Association (JAMA) benchmarks are prominent examples. These frameworks focus heavily on transparency, requiring clear disclosures of authorship, proper attribution of references, currency of the information, and explicit statements regarding financial sponsorships or advertising policies (Silberg et al. [1997], Hirasawa et al. [2012]). In an effort to create a more comprehensive evaluation, some scholars have combined variables from multiple institutional recommendations, such as the HONcode, the Health Information Locator by BIREME-PAHO, Dublin Core metadata standards, and the Web Médica Acreditada, to construct unified "Credibility Indicators". This composite approach was notably used by Guardiola-Wanden-Berghe et al. [2011] to analyze the quality of websites dealing with diets and eating disorders, revealing that the presence of clear authorship and institutional affiliation is a strong predictor of higher informational quality.

Other researchers have augmented these standardized checklists with readability metrics, such as the Flesch Reading Ease score and the Flesch-Kincaid Grade Level, to ensure that the nutritional advice is not only scientifically accurate but also accessible to the general public (Hirasawa et al. [2012]). In these contexts, accuracy is frequently measured by strictly comparing the online claims against aggregated national dietary guidelines (Cardel et al. [2016]).

The shift to social media: modern assessment strategies

While traditional tools like DISCERN and the HONcode remain foundational, they were designed for static web pages and often fall short when applied to the dynamic, highly visual, and brief nature of modern social media platforms. The Web 2.0 environment,

characterized by user-generated content, algorithmic feeds, hidden influencer marketing, and short-form videos, requires updated paradigms. Recognizing this gap, researchers initially resorted to custom methodologies. For example, to evaluate nutritional content on Instagram, Kabata et al. [2022] implemented custom 5-point Likert scales specifically designed to classify posts from *none* to *good quality* based on the verifiability and preparation of the shared knowledge.

More recently, specific rubrics for social networks have emerged. Squires et al. [2023] developed the Social Media Evaluation Checklist to verify the ethical and professional social media practices of registered dietitians and students. It investigates dimensions such as cultural awareness, professionalism, and appropriate financial disclosure. This checklist has proven adaptable; for instance, a modified version was recently applied to evaluate the quality and accuracy of short-form nutritional videos on TikTok, highlighting the platform's severe lack of transparent advertising and evidence-based information (Zeng et al. [2025]).

The PRHISM framework

To systematically address the unique challenges posed by modern digital environments, Denniss et al. [2022] developed the Principles for Health-Related Information on Social Media (PRHISM). Established through a rigorous Delphi study involving experts in health communication, PRHISM represents the current gold standard for evaluating social media health content aimed at non-expert audiences.

Unlike legacy instruments that assume unlimited text length and clear boundaries between content and advertising, PRHISM is intrinsically tailored to the social media landscape. It consists of 13 principles scored on a 5-point Likert scale, categorized into four main themes: accessibility, transparency, authoritative/evidence-based information, and support for the patient-healthcare provider relationship. Crucially, PRHISM introduces criteria that are highly specific to platforms like Instagram, YouTube, and TikTok. For instance, it evaluates *Financial Disclosure* by looking for hidden influencer marketing and paid partnerships. It also assesses the *Action-oriented* nature of the posts, ensuring that messages are succinct and provide sufficient context. Furthermore, PRHISM explicitly evaluates the role of *Images*, demanding that visual elements accurately reflect and do not contradict the written or spoken health messages (Denniss et al. [2022]).

The practical utility of PRHISM has been demonstrated in recent large-scale content analyses. In a comprehensive study of influential Australian Instagram accounts conducted by Denniss et al. [2024], the PRHISM tool was effectively used alongside accuracy assessments to reveal that while some posts might contain factually correct data, their overall quality often remains mediocre due to a lack of references, poor risk-benefit explanations, and missing financial disclosures. Given its comprehensive nature, its specific design for social platforms, and its proven reliability in recent nutritional studies, the PRHISM framework constitutes the methodological cornerstone of the quality evaluation conducted in this thesis.

2.3 Evaluating body-related content

While standardized frameworks such as PRHISM are well-established for assessing the informational quality of health-related content, a significant gap remains in the literature regarding the measurement of appearance-focused content. Specifically, there are currently no consolidated and universally accepted questionnaires designed to systematically evaluate and quantify the level of body-related content within digital videos.

To address similar methodological gaps, recent descriptive content analyses have relied on custom coding schemes. Notably, Munro et al. [2024] conducted a comprehensive investigation of diet culture on TikTok, inductively developing a specific codebook to categorize eating behaviors and body image representations in short-form videos. Their framework identifies key topic-specific variables that are highly indicative of appearance-focused media. These variables include explicit weight measurement, mentions of calories, body checking behaviors, comparisons of the body across different timepoints, and both positive and negative portrayals of body image.

Given the strong alignment between these variables and the visual dynamics of diet-related videos, the framework proposed by Munro et al. [2024] provides an ideal foundation for capturing the nuances of appearance-focused content. Consequently, to systematically quantify the physical appearance focus of YouTube videos, this thesis adapts these specific variables into a structured questionnaire.

2.4 Manual annotation and inter-rater agreement

In digital health research and social media content analysis, the evaluation of subjective or highly contextual variables, such as information quality, scientific accuracy, or specific thematic nuances, relies heavily on rigorous manual annotation procedures. To ensure objectivity and reproducibility, standard methodological practice involves the development of comprehensive coding guidelines and the deployment of multiple independent raters.

A common procedural step is to independently evaluate a subset of the data to test the reliability of the coding scheme before proceeding with the full analysis or automated scaling. For instance, Denniss et al. [2024] independently screened a 10% random sample of nutrition-related Instagram posts using two researchers, resolving any subsequent disagreements through discussion until consensus was reached. Similarly, Zeng et al. [2025] employed three independent researchers to categorize TikTok videos, utilizing consensus discussions to overcome discrepancies. In contexts requiring specialized domain knowledge, experts are often employed as annotators. Syed-Abdul et al. [2013] utilized three independent physicians to classify anorexia-related YouTube videos, bringing in additional reviewers to achieve a majority consensus when initial agreement failed. Furthermore, training annotators through pilot subsets and explicit guidelines is a recurrent and necessary strategy to align human interpretations before actual coding begins (Lai et al. [2022], Ostry et al. [2007]).

To quantify the consistency among independent annotators, researchers rely on various statistical metrics of Inter-Rater Reliability (IRR). Cohen's Kappa is widely adopted for categorical or ordinal data evaluated by two raters (Lai et al. [2022], Ostry et al. [2007]),

while Fleiss' Kappa is applied when three or more raters are involved, yielding moderate agreement levels in studies dealing with complex multimedia content like YouTube videos (Syed-Abdul et al. [2013]). For continuous data or multi-item scales like the DISCERN instrument, the Intraclass Correlation Coefficient (ICC) is frequently employed to capture the degree of consensus (Cruz et al. [2019]).

However, traditional metrics like Cohen's Kappa can be overly sensitive to imbalanced marginal distributions, a known issue referred to as the "prevalence problem". To address this when dealing with highly skewed score distributions, recent studies evaluating medical content, such as the work by Khalil et al. [2025] assessing LLMs as evaluators, have utilized the Brennan-Prediger Kappa, which assumes equal likelihood for all categories under chance agreement and is robust against data skewness. Additionally, when preserving the relative ranking of ordinal scores is more critical than absolute score matching, rank-based metrics like Spearman's rank correlation coefficient (ρ_s) are highly effective.

This established methodological framework directly informs the annotation pipeline adopted in this thesis. This structured manual labeling acts as the fundamental prerequisite for validating the subsequent automated LLM-based annotation strategy.

2.5 Large Language Models as content evaluators and prompting strategies

In recent years, Large Language Models (LLMs) have transcended their original role as mere text generators, proving to be highly effective tools for data annotation and content evaluation. Research has increasingly shown that modern LLMs can act as zero-shot or few-shot evaluators, often matching or even exceeding the reliability and consistency of crowdsourced human annotators (Kalyan [2023]).

Crucially for the context of this thesis, Khalil et al. [2025] demonstrated the robust capability of LLMs in assessing the quality of medical videos on YouTube. By feeding video transcripts into models such as GPT and utilizing the standardized DISCERN questionnaire, the authors showed that LLMs could accurately evaluate complex health information. They employed a guided-scoring prompt design, requiring the models to output specific integer scores (e.g., from 1 to 5) alongside the reasoning for their choices. This study serves as a foundational precedent, validating the methodological feasibility of using GPT models to autonomously answer structured assessment questionnaires based on video transcripts.

The capacity of LLMs to interact with structured clinical and psychometric tools has been confirmed by other recent studies. For instance, Chu et al. [2024] utilized LLMs to probe the collective mindset of online Eating Disorder (ED) communities on social media. By aligning the Llama-3 model to the specific language of these communities via instruction tuning, the authors successfully administered the SWED (Screening for Weight and Eating Disorders) questionnaire directly to the model, effectively revealing the risk levels regarding unhealthy diet and body concerns in digital spaces. Similarly, Nori et al. [2023] showcased GPT's exceptional zero-shot performance on the United States Medical Licensing Examination (USMLE). Their findings proved the model's inherent capacity to comprehend complex health-related multiple-choice questions without the need for

specialized fine-tuning, highlighting the strong baseline calibration of GPT in medical reasoning.

The role of prompting techniques

The success of LLMs in structured evaluation tasks heavily depends on the adopted prompting strategies. As noted by La Rocca [2025] in a study detecting conspiracy theories on YouTube, while LLMs generally achieve high recall, their precision is highly sensitive to prompt formulation. The author found that providing formal, explicit definitions of the criteria within the prompt (definition-based prompting) significantly enhanced the classification performance, particularly for larger models. This aligns with findings by Parikh et al. [2023], who demonstrated that using detailed intent descriptions allows LLMs to perform highly competitive zero-shot classification, bypassing the need for extensive training data or fine-tuning.

To fully contextualize this advancement, it is essential to distinguish between zero-shot and few-shot prompting. In a zero-shot scenario, the model receives solely a task description or formal definitions, relying entirely on its pre-trained knowledge to infer the output (Parikh et al. [2023]). This approach is highly scalable and eliminates the bottleneck of collecting training data.

Conversely, few-shot prompting prepends a small number of curated demonstration examples to the query to guide the model through in-context learning (Brown et al. [2020]). While few-shot prompting consumes more context tokens, it implicitly teaches the model the desired output format and evaluation boundaries through pattern recognition.

To tackle tasks requiring deeper cognitive processing and evaluation, the Natural Language Processing community has shifted toward reasoning-eliciting techniques. The Chain-of-Thought (CoT) prompting strategy, introduced by Wei et al. [2022], significantly improves performance by forcing the model to generate intermediate, step-by-step reasoning before outputting a final answer. While originally utilizing few-shot exemplars to demonstrate this reasoning process, Kojima et al. [2022] proved that LLMs can act as excellent zero-shot reasoners simply by appending an instruction like *Let's think step by step* to the prompt. The effectiveness of CoT has proven particularly valuable in social media analysis; for instance, Zhang et al. [2024] applied a step-by-step question-answering approach to perform zero-shot stance detection on social media platforms, demonstrating state-of-the-art performance and highlighting that semantic-level reasoning prompts generally outperform purely word-level instructions.

The application of CoT is particularly beneficial in evaluation and Natural Language Understanding (NLU) contexts. Zhong et al. [2023] compared ChatGPT with fine-tuned BERT models on the GLUE benchmark, observing that while standard zero-shot LLMs sometimes struggle with nuanced semantic textual similarity or negative paraphrasing, the integration of manual few-shot CoT prompting drastically narrows this performance gap.

In the specific context of clinical language understanding, Wang et al. [2023b] proposed Self-Questioning Prompting (SQP), a novel strategy where the model is instructed to ask itself clarifying questions about the medical scenario and answer them internally before

generating the final classification. This mimics the human reasoning process during complex medical evaluations and has been shown to outperform standard CoT in healthcare tasks. Additionally, to mitigate the stochasticity of greedy decoding, techniques like Self-Consistency (Wang et al. [2023a]) have been introduced. Instead of taking the single most likely response, self-consistency samples multiple diverse reasoning paths from the model and aggregates the answers through a majority vote, leading to highly robust predictions.

To further push the boundaries of complex reasoning, researchers have proposed progressive prompting frameworks. Zhou et al. [2023] introduced Least-to-Most Prompting, a strategy that explicitly decomposes a complex problem into a sequence of simpler subproblems, solving each sequentially so that the answer to a previous subproblem facilitates the next. Adapting this philosophy specifically for text classification, Sun et al. [2023] proposed Clue And Reasoning Prompting (CARP). Instead of a generic reasoning step, CARP forces the LLM to first extract superficial evidence (e.g., keywords, semantic relations) from the text, use these clues as premises for a diagnostic reasoning process, and finally output the classification decision. This structured progression heavily mimics human decision-making and drastically improves accuracy on complex linguistic phenomena compared to standard CoT .

However, as highlighted by Wu et al. [2025], employing advanced reasoning frameworks like Chain-of-Thought (CoT) introduces significant practical trade-offs when processing large-scale datasets. While CoT enhances zero-shot accuracy, it can paradoxically degrade few-shot performance due to token overload, increased prompt complexity, and the model's tendency to overfit the provided examples. Consequently, simpler and more scalable approaches, such as standard zero-shot prompting, often achieve highly competitive results while drastically reducing API costs and computational time. This balance between analytical depth and efficiency directly justifies the practical decision in this study to adopt a standard zero-shot configuration for the massive annotation of video transcripts.

Chapter 3

Method

To address the research questions, it is first necessary to collect data for each YouTube video deemed relevant to the study. This includes not only metadata, but also titles, descriptions, and transcripts, as detailed in section 3.1. The subsequent data pre-processing phase is described in section 3.2, and includes procedures such as duplicate removal and the handling of missing values. This stage is followed by a preliminary exploratory analysis, which provided an overall overview of the collected data.

Semantic data play a central role in investigating video topics, information quality, and body-related content. Topic extraction is performed through non-negative matrix factorization (NMF), as described in section 3.4. This approach provides a comprehensive overview of the main themes discussed on social media under the broad umbrella of diet and weight-loss content, thereby addressing the first research question (Q1).

To evaluate information quality and body-related content, two dedicated scores are adopted. These are derived from validated questionnaires (3.6) and operationalized through GPT-based prompting procedures (3.8). Together with the extracted topics and the collected metadata, these scores form the basis for addressing the second research question (Q2) through univariate analyses and linear regression models. Similar univariate and regression analyses are then conducted to answer the third research question (Q3), which focuses on user engagement.

Finally, the fourth research question (Q4) requires the development of a standard neural network model designed to predict the quality and body-related scores based on topics and metadata. While conceptually related to Q2, this step adopts a predictive rather than exploratory perspective.

A detailed description of all analytical steps undertaken to address the research questions is provided in section 3.9, while an overview of the statistical tools employed throughout the study can be found in Appendix A.

3.1 Data collection

We carried out data collection through two separate procedures, and the results were saved in both `.json` and `.tsv` formats. The `.json` files store the raw data exactly as

returned by the YouTube API (see subsection 3.1.1), preserving the full structure of the original response. The `.tsv` files, instead, contain the processed and structured data, making them more suitable for subsequent analysis.

Dynamic collection

A first collection began on January 8, 2025 on behalf of the ISI Foundation researcher Yelena Mejova, and is still ongoing. Based on this dynamic collection, each day the most relevant videos uploaded the day before on the topic of interest are chosen (*initial data*), and these same videos are periodically retrieved to document the evolution of their statistics as well as any modifications made by the creator (*periodic data*).

Initially, only the video metadata (title, description, channel, engagement statistics, etc.) are collected, while *periodic data* include video metadata, channel metadata, and a sample of video comments. Periodic checks are performed every day during the first week after the video is uploaded, then every week until 3 months from the upload. If, at the end of this three-month period, the last two crawls result in different engagement numbers (likes or comments), the periodic collection continues on a weekly basis until this condition is no longer satisfied.

Retrospective collection

A second collection (*popular data*) was instead conducted by the author between November 10 and December 1, 2025, retrospectively searching for the most popular videos uploaded within the time period between January 8, 2025 and June 7, 2025. We implemented this procedure to complete the existing dataset and to ensure that information about the videos with the highest levels of user interaction was not lost. Collecting these data is essential for understanding the potential reach of risky content, as it provides insight into which types of videos on the topic users are most likely to watch.

To perform the data collection, we first divided the five-month period into one-day windows, and for each window the videos uploaded on that day with the highest view counts were selected. The same procedure was then repeated using one-month windows and ten-day windows. We obtained the final set of results by taking the union of these three sub-collections.

For each video, we collected its metadata, the metadata of its channel, and a sample of its comments as they appear at the time of retrieval.

3.1.1 API queries

Both collections relied on the YouTube Data API v3 (API reference), the official YouTube Application Programming Interface that allows access to public platform content and can be easily integrated into Python scripts through the `googleapiclient.discovery` package by specifying the service as `"youtube"` and the version as `"v3"`. Once we create the client object using a personal API key, the querying process relies on the methods described below, which were used in both data collections.

Platform search and video IDs retrieval

The `search().list()` method allows, in the case of this study, to perform a YouTube search using specific keywords and to obtain the IDs of the corresponding videos. Since the research interest focuses on content related to dieting and weight loss, we used the following keywords for both collections: *diet, dieting, diets, weight loss, weightloss, fasting, nutrition*. We combined these keywords in an alternative (OR) logic in order to maximize the number of relevant results. Their selection was informed by previous studies conducted by ISI Foundation researchers, particularly the work of Mejova and Suarez-Lledó [2020].

We set the parameters to retrieve videos that are accessible in the United States and have English as their relevance language. We set the order of search results to `relevance` for the dynamic collection and `viewCount` for the retrospective one. The exact internal algorithms selecting the videos are proprietary and running the same query multiple times can lead to different results.

The output of the method, once executed using `execute()`, is a dictionary representing a results page. It contains, under the key `"nextPageToken"`, the identifier required to access the subsequent page by executing another `search` query. In addition to this identifier, the dictionary includes the total number of results across all pages (`pageInfo.totalResults`), the number of results per page (`pageInfo.resultsPerPage`), and under the key `"items"`, a list of dictionaries each containing the essential metadata of the videos on the current page, including their IDs.

For some inexplicable reasons, after a few iterations of the page token mechanism, the search endpoint returns a response dictionary that lacks the `"nextPageToken"` key, even though the total number of results has not been reached yet, not even remotely. Due to this behavior, it is impossible to set an a priori number of IDs to be collected, but only an upper bound.

Retrieval of video metadata

After obtaining video IDs, the second step involves downloading complete metadata for each video, through the API's `videos().list()` endpoint. We grouped and passed IDs to the method in batches of 50 (the API maximum per request), ensuring efficient retrieval, reduced API load, and safe intermediate storage.

Retrieval of channel metadata

Channel-level metadata was retrieved using the `channels().list()` method. As with videos, channel IDs were deduced from video metadata and processed in batches of 50.

Retrieval of comment threads

To retrieve user comments associated with each video, we uses the `commentThreads().list()` endpoint of the YouTube Data API. For every video ID in the dataset, the code attempts to download up to 1000 comments by iteratively querying the API with a page token mechanism. Each request returns comment threads in reverse chronological order (`order = "time"`), including both top-level comments and replies

when available. We stored the raw `.json` responses and the structured `.tsv` data in compressed form.

3.1.2 Transcripts extraction

Video transcripts cannot be collected through the YouTube Data API v3, but are essential for accurately selecting videos and identifying video topics. Therefore, we used another YouTube API, the YouTube Transcript Api, designed to extract both manually created and automatically generated captions, based to video IDs.

Due to restrictions on transcript access imposed by YouTube, only a limited number of video transcripts (on the order of tens) can be retrieved from a single IP address within a given time frame, making the collection process very time-consuming.

3.1.3 Data selection

After data collection, we immediately applied some selection criteria, and only videos meeting the following conditions are kept:

- The default audio language must be english;
- The video duration must be between one minute and one hour;
- There must be at least one keyword match in both the title and the transcript. If there is no match in the title, there must be at least three matches in the transcript.

As a result of the selection criteria, transcripts are essential, and a missing transcript leads to the automatic removal of the video's data from the dataset. The irregular restrictions on transcript retrieval caused massive video deletions during certain periods of the dynamic data collection. This phenomenon can be observed by examining the number of videos selected each day based on the conditions above, as shown in Figure 3.1 and Figure 3.2.

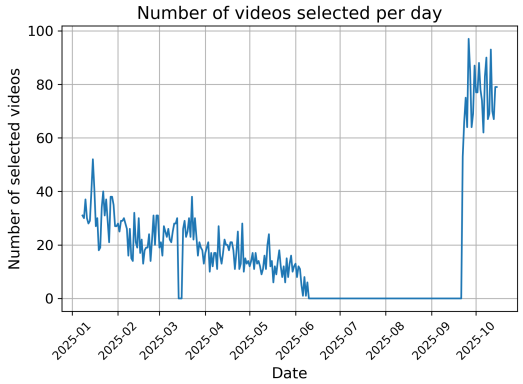


Figure 3.1: Number of videos (*initial data*) selected each day during the dynamic collection.

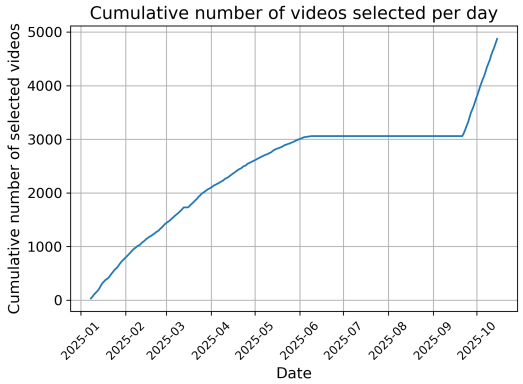


Figure 3.2: Cumulative number of videos (*initial data*) selected each day during the dynamic collection.

Table 3.1: Video metadata: `videos().list()` endpoint variables details. Descriptions from the API reference.

Name	Type	Description	Example
id	string	The ID that YouTube uses to uniquely identify the video.	9hfKIZ3j1Kg
snippet.title	string	The video's title. The property value has a maximum length of 100 characters and may contain all valid UTF-8 characters except <and>.	How Losing Weight Changed my Life | Manage your Weight Loss Expectations to Find Success in 2025
snippet.channelID	string	The ID that YouTube uses to uniquely identify the channel that the video was uploaded to.	UClxxh1umeWjxc6w9lKstdFw
snippet.channelTitle	string	Channel title for the channel that the video belongs to.	Janeé
snippet.tags	list	A list of keyword tags associated with the video. Tags may contain spaces. The property value has a maximum length of 500 characters.	['top weight loss tips', 'best weight loss tips', 'best weight loss tips for women', 'best weight loss tips and tricks',...]
snippet.description	string	The video's description. The property value has a maximum length of 5000 bytes and may contain all valid UTF-8 characters except <and>.	I have lost 124 pounds from my highest weight and in today's videos, I am going to share with you how losing weight and having [...]
snippet.categoryId	int	The number of the YouTube video category associated with the video.	22
snippet. defaultLanguage	string	The language of the text in the video's title and description.	en-US
topicDetails. topicCategories	list	A list of Wikipedia URLs that provide a high-level description of the video's content.	['https://en.wikipedia.org/wiki/Health', 'https://en.wikipedia.org/wiki/Lifestyle']
statistics.viewCount	unsigned long	The number of times the video has been viewed.	175
statistics.likeCount	unsigned long	The number of users who have indicated that they liked the video.	24
statistics. commentCount	unsigned long	The number of comments for the video.	11

From June 10 to September 21, 2025, no videos were selected. Therefore, we considered only the five-month period from January 8 to June 7, 2025, for all subsequent analyses, and we adopted the same time frame for the retrospective data collection.

In addition to reducing the number of videos considered in the dataset, we also reduced the number of variables. From the full set of variables retrieved through the different API queries, we selected a smaller subset based on the research questions. The chosen variables for each query type are summarized in Table 3.1 (video metadata), Table 3.2

Table 3.2: Channel metadata: `channels().list()` endpoint variables details. Descriptions from the API reference.

Name	Type	Description	Example
id	string	The ID that YouTube uses to uniquely identify the channel.	UClxxh1umeWjxc6w9lKstdFw
snippet.title	string	The channel's title.	Janeé
snippet.description	string	The channel's description. The property's value has a maximum length of 1000 characters.	Hey Babes! I am so happy you're here! Welcome to this safe space for wellness, fitness, self-care, personal development, life lessons, [...]
snippet.customUrl	string	The channel's custom URL.	@janeeleesanders
snippet.publishedAt	date object	The date and time that the channel was created. The value is specified in ISO 8601 format.	2020-11-10T20:24:10.08703Z
snippet.country	string	The country with which the channel is associated.	US
statistics.viewCount	unsigned long	The sum of the number of times all the videos in all formats have been viewed for a channel.	8232517
statistics. subscriberCount	unsigned long	The number of subscribers that the channel has. This value is rounded down to three significant figures.	17100
statistics. videoCount	unsigned long	The number of public videos uploaded to the channel. Note that the value reflects the count of the channel's public videos only, even to owners. This behavior is consistent with counts shown on the YouTube website.	1634
topicDetails. topicCategories	list	A list of Wikipedia URLs that describe the channel's content.	['https://en.wikipedia.org/wiki/Lifestyle', 'https://en.wikipedia.org/wiki/Health', 'https://en.wikipedia.org/wiki/Physical_fitness']
brandingSettings. channel.keywords	string	Keywords associated with the channel. The value is a space-separated list of strings. Channel keywords might be truncated if they exceed the maximum allowed length of 500 characters or if they contained unescaped quotation marks (").	"janee lee sanders" wellness "wellness girlie" "wellness journey" fitness "at home workouts" "gym workouts" workouts nutrition "weight loss" [...]

(channel metadata), and Table 3.3 (comment threads), together with the data type, a brief description (taken from the API reference), and an example based on a randomly sampled record from the early rows of the dataset.

We also extracted the variables `snippet.defaultAudioLanguage` and `contentDetails.duration` during the *initial data* and *popular data* collections to apply the selection criteria described above. We then removed them, and stored the video

Table 3.3: Comment threads: `commentThreads().list()` endpoint variables details. Descriptions from the API reference. `sts` stands for `snippet.topLevelComment.snippet`.

Name	Type	Description	Example
id	string	The ID that YouTube uses to uniquely identify the comment thread.	UgwfVdK9hJ-E-2EsWF54AaABAg
snippet.channelId	string	The YouTube channel that is associated with the comments in the thread.	UClxxh1umeWjxc6w9lKstdFw
snippet.videoId	string	The ID of the video to which the comments refer.	9hfKIZ3j1Kg
sts.authorDisplayName	string	The display name of the user who posted the comment.	@tbowman85
sts.authorChannelId.value	string	The ID of the comment author's YouTube channel, if available.	UCscwT0cPLR-lV1QILYzG0oA
sts.textDisplay	string	The comment's text. The text can be retrieved in either plain text or HTML. Even the plain text may differ from the original comment text. For example, it may replace video links with video titles.	I'm always cold now. I had rny gastric bypass ...
sts.likeCount	unsigned integer	The total number of likes (positive ratings) the comment has received.	1
sts.publishedAt	date object	The date and time when the comment was orignally published. The value is specified in ISO 8601 format.	2025-01-08T04:56:22Z
sts.updatedAt	date object	The date and time when the comment was last updated. The value is specified in ISO 8601 format.	2025-01-08T04:56:22Z
snippet.totalReplyCount	unsigned integer	The total number of replies that have been submitted in response to the top-level comment.	1

duration as a number of seconds in the variable `duration_secs`. We also added the `transcript` variable to the dataset.

Since video metadata, channel metadata, and comment threads are collected periodically for each video during the *periodic data* collection, every channel query is associated with a video ID. Moreover, each query of any of these three types is linked to the date on which the video was first retrieved (`searchdate` or `search_date`), the date on which the periodic check was performed (`periodic_date` or `check_date`), and the number of days elapsed between the upload date and the check (`days_after`).

3.2 Data pre-processing

Once all the data had been collected, selected, and stored in a structured format, we carried out a pre-processing phase to ensure data cleaning and completeness.

3.2.1 Initial and popular data

For the *initial data* (initial video metadata of the dynamic collection) and the *popular data* (from the retrospective collection), we followed the following steps:

1. Drop records whose fields were incorrectly split, characterized by null values in `searchdate`.
2. Drop duplicated records. Some videos appeared in the search results on two consecutive days, for reasons that remain unclear. In these cases, only the first occurrence was retained.
3. Map category IDs to their corresponding category names using a conversion dictionary.
4. Convert `searchdate` field from float to string.
5. Drop records with missing (null) like counts or missing (null) comment counts. All videos have non-missing view counts.

The number of *initial data* records selected was 3052, which was reduced to 3048 after step 1 and to 3042 after duplicate removal (step 2). In fact, 6 videos had been retrieved on consecutive days. However, the largest reduction occurred in step 5, which brought the number of records down to 2870, a set of data that will be denoted as *initial data clean*.

After *periodic data* cleaning (see following subsection), we performed an additional step on a copy of the *initial data clean* dataset, retaining only the data of the 1157 videos that had been consistently retrieved in each of the 20 mandatory periodic checks. This subset of records is referred to as *initial data with complete checks*.

Concerning the *popular data*, after filtering and pre-processing, the final dataset comprises 384 videos, which is substantially lower than the 1,952 videos initially retrieved through the search procedure.

3.2.2 Periodic data

Because some videos appeared in the search results on two consecutive days, each of them was retrieved twice (on two consecutive days) during every periodic check. Since we retained the first occurrence in the *initial data*, we applied the same criterion to all periodic checks of those videos, both for periodic video metadata and periodic channel metadata.

During the pre-processing of periodic video metadata, we also removed records with missing statistics, and mapped category IDs to their corresponding names, following the same procedure described in steps 3 and 5 of the *initial data* pre-processing.

We examined more closely the records with missing statistics in the periodic channel metadata. In fact, no channel was missing either subscriber or video counts, and only three channels were missing the view count in a single periodic check. For each of these three channels, the view count was identical in the checks immediately before and after the gap, so we filled the missing value using these neighboring values.

After this processing, we obtained 43,907 records in what we call *periodic video metadata clean* and 40,006 records in *periodic channel metadata clean* (together: *periodic data clean*).
The subset of records in the periodic video metadata corresponding to videos that were consistently retrieved in each of the 20 mandatory periodic checks is referred to as *periodic video metadata with complete checks*. All the video metadata collected in the periodic checks of these videos are present, although channel metadata may still be missing for some checks. This is disregarded, as we are not focusing on the temporal evolution of channel data.

Regarding the comment thread data, we performed no pre-processing, as we did not included these data in the analyses conducted in this study. We nevertheless retained comment thread data in the dataset to ensure greater completeness and to support potential future research. This allows subsequent studies to process and analyze such data according to their specific objectives, without being constrained by the pre-processing choices adopted here.

3.3 Preliminary data exploration

Following data collection and pre-processing, we conducted a preliminary exploratory phase. During this stage, we computed descriptive statistics on the collected data (see section 4.1). In addition, we reviewed a subset of YouTube videos, selected either randomly or based on insights emerging from the descriptive analysis, in order to gain a broader understanding of the content under investigation and the key phenomena characterizing it.

A following phase of the data exploration involved extracting keywords from the text generated for each video by concatenating the title, description, and transcript, and then converting the result to lowercase.

To perform the keyword extraction, we used a python library based on YAKE! (Yet Another Keyword Extractor) unsupervised and domain-independent algorithm. YAKE! automatically extracts keywords directly from the text, without need of context. Its core idea is that the most informative words in a text tend to show distinctive statistical patterns compared to the rest of the vocabulary. To capture this, YAKE! computes, for each word, a set of local features that describe how it behaves within the document. These features include how often the word appears, where it tends to occur, how evenly it is distributed throughout the text, how it co-occurs with neighboring terms, and how relevant it is in combination with the words around it. All these aspects are combined into a single score that highlights words with moderate frequency, relatively uniform distribution, and meaningful co-occurrences.

After assigning a score to each word, YAKE! creates keyword candidates by grouping adjacent words and computing a score for each two-word expression based on its component terms. The algorithm also penalizes redundant combinations and finally selects the 5 expressions with the lowest scores (lower values indicating higher relevance).

Since all features are extracted from each text independently and no external resources or training data are used, YAKE! tends to identify several trivial keywords that are

common across many videos and do not contribute much to topic discrimination. On the other hand, it remains fast and lightweight.

3.4 Topic modeling

We used semantic information, coming from data collection and pre-processing, to address the first research question (Q1), which concerns topic identification. Specifically, the objective is to characterize the themes discussed on social media, particularly YouTube, within the broad domain of diet and weight loss.

The literature suggests that the most commonly adopted topic modeling techniques in similar contexts are BERTopic, Latent Dirichlet Allocation (LDA), and Non-negative Matrix Factorization (NMF) (see section 2.1). LDA is widely used, especially when dealing with relatively structured textual data such as news articles or scientific documents. However, its performance may be less satisfactory in highly unstructured settings, such as automatically generated video transcripts, or when the document collection is strongly mono-thematic.

We initially tested BERTopic on the transcript data. Although it identified a reasonable number of topics, a large proportion of the documents were classified as outliers. Reducing the number of outliers required increasing the number of topics substantially, resulting in overly specific and fragmented themes.

NMF, on the other hand, has been successfully applied in previous studies with unstructured data and we adopted Cheng et al. [2022] as reference paper. Adopting a similar methodological framework proved effective in our context as well. The resulting topics were coherent and interpretable, and their validity was further supported by the manual evaluation conducted by the annotators (see section 3.7 and section 4.2).

For these reasons, we did not consider additional topic modeling techniques.

3.4.1 Non-negative Matrix Factorization for topic modeling

Non-negative Matrix Factorization (NMF) is a dimensionality reduction technique that decomposes a non-negative matrix into the product of two lower-rank non-negative matrices. Given a document–term matrix $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n \times m}$, where n denotes the number of documents and m the number of terms, NMF seeks an approximate factorization of the form:

$$\mathbf{X} \approx \mathbf{WH},$$

where $\mathbf{W} \in \mathbb{R}_{\geq 0}^{n \times k}$ and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{k \times m}$, with k representing the number of latent topics. The matrix $\mathbf{W}$ encodes the representation of each document in terms of the k topics, while $\mathbf{H}$ captures the contribution of each term to each topic.

The factorization is obtained by minimizing a divergence measure between $\mathbf{X}$ and $\mathbf{WH}$, typically the Frobenius norm:

$$\min_{\mathbf{W},\mathbf{H} \geq 0} \|\mathbf{X} - \mathbf{WH}\|_F^2,$$

subject to non-negativity constraints. The non-negativity property makes NMF particularly suitable for topic modeling in textual data.

The value of this objective function is referred to as the *reconstruction error*, as it quantifies how accurately the product **WH** approximates the original matrix **X**.

To apply NMF, we first construct the document–term matrix $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n \times m}$ summarizing the textual data associated with the videos in our dataset. Each document corresponds to the concatenation of a video's title, description, and transcript, forming the rows of **X**. The columns represent (most of) the terms appearing at least once across the entire corpus, and the entries are defined according to the vectorization procedure described later.

Each entry of the resulting matrix **W**, as mentioned above, is a non-negative value indicating the extent to which a given video is characterized by a particular topic. Consequently, each topic can be treated as a continuous variable within the dataset.

However, we also considered an alternative representation in which topics are modeled as categorical features, allowing each video to be associated with zero, one, or multiple topics. To this end, we defined a threshold for each topic, and assigned only videos with weights exceeding this threshold to the corresponding topic. Following the reference paper, we set the threshold for each topic equal to the mean of its weights (i.e., the values in the corresponding column of **W**) plus one standard deviation.

Both representations jointly contribute to addressing the first research question (Q1).

3.4.2 Text pre-processing for topic modeling

Since NMF operates on single-word frequencies without accounting for word order or context, we applied a dedicated pre-processing pipeline to remove uninformative words and special characters.
The pre-processing applied to titles and descriptions consisted of the following steps:

1. Replacement of escape sequences with a single space character;
2. Conversion to lowercase;
3. Removal of URLs;
4. Removal of emojis;
5. Removal of commas in numbers greater than or equal to 1,000;
6. Removal of tags (preceded by #) and mentions (preceded by @);
7. Replacement of multiple consecutive spaces with a single space character.

For transcripts, we performed the following steps:

1. Replacement of escape sequences with a single space character;
2. Removal of text enclosed in square brackets (e.g., [music]);
3. Removal of timestamps (e.g., 00:01:23);
4. Removal of caption tags (e.g., <b>);
5. Conversion to lowercase;
6. Removal of commas in numbers greater than or equal to 1,000;
7. Removal of common filler words;

8. Removal of consecutively duplicated words;
9. Removal of punctuation surrounded by spaces;
10. Replacement of multiple consecutive spaces with a single space character.

By *filler words*, we refer to common disfluencies in spoken language that do not convey informative semantic content. Their removal is a common practice in natural language processing to improve model performance and reduce noise. We defined the list of filler words based on linguistic references (Tree [1995]), previous literature (Cheng et al. [2022], Lalk et al. [2024], Lai et al. [2022]), and an exploratory analysis of the collected transcripts. The filler words considered at this stage are:

> *um, uh, erm, hmm, mm, yeah, huh, huhm, eh, ah, basically, I mean, okay*

We then concatenated title, description, and transcript using ". " as a separator. We subsequently applied the following more aggressive pre-processing steps to the resulting text, hereafter referred to as *text*:

1. Removal of numbers;
2. Removal of punctuation;
3. Tokenization;
4. Removal of stop words using the `stopwords` module from `nltk.corpus`;
5. More extensive removal of filler words;
6. Removal of discourse adverbs;
7. Removal of words shorter than three characters;
8. Stemming;
9. Re-concatenation of tokens into a single string.

During these more aggressive steps, the filler words considered are:

> *uh, um, yeah, okay, ok, like, youknow, gonna, wanna, gotta, cuz, oh, ah, eh, hmm, guys, stuff, thing, things, mean, actually, basically, literally*

Discourse adverbs organize discourse by signaling logical relations, contrast, emphasis, or speaker attitude rather than contributing core propositional meaning (Fraser [1990]). The ones removed include:

> *really, very, just, maybe, probably, definitely, honestly, seriously, kinda, sorta, almost, totally, simply, pretty*

Finally, stop words are high-frequency function words (e.g., *the, and, is, of*) that typically carry little semantic meaning and are commonly removed during text pre-processing to reduce noise and dimensionality. In this study, we identified stop words using a standard Python library.

3.4.3 Vectorizer application

An important component of the pipeline is the choice of the vectorizer used to transform each pre-processed text into a numerical representation, i.e., to define the entries of the document–term matrix $\mathbf{X}$.

Following the study conducted by Cheng et al. [2022], we adopted the Term Frequency–Inverse Document Frequency (TF–IDF) weighting scheme. TF–IDF measures the importance of a term within a document relative to the entire corpus: it increases proportionally to the term frequency in the document, and decreases with the term frequency across the corpus, thus penalizing common but non-discriminative words.

The TF–IDF weight of term t in document d is defined as:

$$\text{tfidf}(t, d) = \text{tf}(t, d) \cdot \text{idf}(t),$$

where the term frequency is

$$\text{tf}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

and the inverse document frequency is

$$\text{idf}(t) = \log\left(\frac{n}{\text{df}(t)}\right).$$

Here, $f_{t,d}$ denotes the raw frequency of term t in document d, $\text{df}(t)$ is the number of documents containing term t, and n is the total number of documents in the corpus.

In our application, we adopted the smoothed TF–IDF variant, where the inverse document frequency is defined as:

$$\text{idf}(t) = \log\left(\frac{n+1}{\text{df}(t)+1}\right) + 1.$$

The smoothing prevents division by zero and mitigates the impact of terms appearing in nearly all documents.

We introduced two additional thresholds: *max_df* and *min_df*. The parameter *max_df* represents an upper bound on document frequency (expressed as a proportion). Terms whose document frequency exceeds this threshold are treated as corpus-specific stop words and removed. This parameter is considered a hyperparameter and is varied over the following values, selected after several exploratory attempts:

$$max\_df \in \{0.30, 0.325, 0.35, 0.375, 0.40, 0.45, 0.50\}.$$

Conversely, *min_df* defines a lower bound on document frequency: terms appearing in fewer than 5% of the documents are discarded, as they are considered too specific. In our study, we fixed *min_df* at 0.05, as suggested by our reference literature paper for NMF.

After computing the TF–IDF weights for each selected term in each document, the resulting values populate the document–term matrix $\mathbf{X}$.

3.4.4 Number of topics

Another crucial parameter is the number of topics k, which determines the dimensions of the factor matrices $\mathbf{W}$ and $\mathbf{H}$.

For each value of the hyperparameter *max_df*, we tested all integer values of k in the range $6 \leq k \leq 20$. We chose this range to balance interpretability and specificity, excluding overly generic or excessively granular topic structures.

For each combination of *max_df* and k, we computed both the context vector–based topic coherence (see below) and the reconstruction error. Since the reconstruction error did not provide meaningful insights, model selection was primarily based on coherence.

For each value of *max_df*, we retained the two values of k yielding the highest coherence, resulting in $7 \times 2 = 14$ candidate models. We then conducted a final manual inspection to select the most interpretable model, based on the top 10 words associated with each topic.

Context vector–based topic coherence

The c_v coherence measure (context vector–based topic coherence), introduced by Röder et al. [2015], evaluates the semantic consistency of a topic by combining boolean sliding-window co-occurrence statistics with a vector-based similarity framework.

A boolean sliding window of fixed length $L = 110$ tokens is moved over the reference corpus (in our case, the collection of all *texts*) advancing one token at a time. Each window defines a "virtual document" d'. Let D denote the total number of such virtual documents, and let $W = \{w_1, \ldots, w_M\}$ be the set of the top $M = 10$ words associated with a given topic. These correspond to the terms with the highest weights in the respective row of the matrix $\mathbf{H}$, i.e., the terms that contribute most strongly to that topic. Word probabilities are estimated using boolean document frequencies over the virtual documents:

$$P(w_i) = \frac{|\{d' : w_i \in d'\}|}{D}, \qquad P(w_i, w_j) = \frac{|\{d' : w_i, w_j \in d'\}|}{D}.$$

For each pair (w_i, w_j), the normalized pointwise mutual information (NPMI) is computed as:

$$\mathrm{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}.$$

Each word $w_i \in W$ is represented by a context vector

$$\mathbf{v}_{w_i} = \big(\mathrm{NPMI}(w_i, w_j)\big)_{j=1}^{M},$$

which encodes its semantic association with the other topic words.

Under the $S_{\text{one-set}}$ segmentation, each word w_i is compared with the full topic word set W. The confirmation measure is computed as the cosine similarity between the corresponding context vectors:

$$\tilde{m}_{\cos}(w_i, W) = \cos(\mathbf{v}_{w_i}, \mathbf{v}_W),$$

where

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|_2 \, \|\mathbf{v}\|_2}, \qquad \mathbf{v}_W = \sum_{k=1}^{M} \mathbf{v}_{w_k}.$$

The coherence of the single topic is obtained by averaging these confirmation measures (aggregation):

$$c_v(W) = \frac{1}{M} \sum_{i=1}^{M} \tilde{m}_{\cos}(w_i, W).$$

The overall model coherence is finally computed as the arithmetic mean of the coherence scores across all topics. Higher values indicate greater semantic relatedness among topic words and thus higher interpretability.

3.5 Non-semantic variables

Titles, descriptions, transcripts, and extracted topics provide valuable insights into the semantic content of a video. Equally important, however, are non-semantic metadata, which can offer complementary information beyond the textual content itself.

Elements such as channel age, title length, or the number of external URLs included in the description may vary substantially across creators and can offer additional explanatory power in understanding differences in content production and communication strategies. We refer to these variables, together with other non-semantic features related channels and *texts*, as *non-semantic variables*.

If meaningful relationships can be identified between these variables and either information quality or the level of body-related content, they offer a key advantage: they are deterministic and computationally inexpensive, while also being useful for gaining insights into, or predicting, potential risk.

Most of the variables are inspired by the study conducted by Cossard et al. [2020], while others were identified as relevant based on the descriptive statistics (4.1) computed during the initial exploratory phase.

The variables can be grouped into four main categories, which are listed below.

General video characteristics

- Video duration (seconds) – `duration_secs`;
- Video category retrieved via the API – `categoryName`;
- Video duration normalized by the category median – `duration_norm`.

General channel characteristics (last available snapshot)

- Number of subscribers – `channelSubscriberCount`;
- Number of videos – `channelVideoCount`;

- Channel age (days, computed with respect to November 30) – `channelAge`;
- Country – `channelCountry`;
- Channel description length (word count) – `channelDescription_len`;
- Channel description length (character count) – `channelDescription_lenChar`;
- Topic category – `channelCategories`;
- Upload frequency (videos per day) – `channel_frequency`.

Textual content length features

- Transcript length (word count) – `transcript_len`;
- Description length (word count) – `description_len`;
- Description length (character count) – `description_lenChar`;
- Title length (character count) – `title_lenChar`.

Textual content characteristics

- Average word length across all textual data – `text_wordLen`;
- Number of uppercase letters in title and description – `title_desc_uppercase`;
- Ratio of uppercase letters to total letters in title and description – `title_desc_uppercaseRatio`;
- Number of emojis in title and description – `title_desc_emoji`;
- Ratio of emojis to total characters in title and description – `title_desc_emojiRatio`;
- Number of exclamation marks in title and description – `title_desc_exlam`;
- Ratio of exclamation marks to total characters in title and description – `title_desc_exlamRatio`;
- Number of external links (URLs) in the description – `description_links`;
- Number of hashtags (preceded by #) in the description – `description_hashtags`;
- Number of mentions (preceded by @) in the description – `description_mentions`;
- Lexical diversity of the transcript (unique words / total words) – `transcript_lexicalDiversity`.

The number of variables that we considered is relatively high; however, substantial correlations are expected among them. In particular, strong correlations are likely between word-based and character-based measures, as well as between composite variables and their underlying components, such as the ratio of uppercase letters and the absolute number of uppercase letters or total character count.

It is important to note that counting letters is not equivalent to counting characters, as characters also include punctuation marks, emojis, and other special characters.

3.5.1 Variable selection

To reduce redundancy and mitigate multicollinearity in the linear regression models (A.2), we conducted an initial screening of the variables based on the previously discussed relationships and the inspection of the correlation matrix.

We removed the composite variables `duration_norm`, `channel_frequency`, and `transcript_lexicalDiversity`. In contrast, for stylistic features such as uppercase letters, emojis, and exclamation marks, we preferred composite indicators over their underlying raw components in order to ensure independence from title and description lengths. This choice led to the exclusion of `title_desc_uppercase`, `title_desc_emoji`, and `title_desc_exlam`.

As expected, a strong correlation emerged between video duration and transcript length (measured in words), which motivated the removal of `transcript_len`. Finally, we retained word-based length measures for video and channel descriptions in place of character-based measures, resulting in the exclusion of `description_lenChar` and `channelDescription_lenChar`.

3.6 Information quality and body-related content: the questionnaires

Building on the previously extracted semantic and non-semantic features, the next step is to relate these characteristics to the notion of risk. This requires translating the concept of risk into measurable and systematically analyzable dimensions.

Low information quality and a high level of body-related content in YouTube videos are key contributors to highly risky content. A major challenge in this context lies in the measurability of these dimensions. Based on previous literature (see section 2.2), the methodological choice is therefore to adopt questionnaire-based approaches.

3.6.1 Quality questionnaire

Several questionnaires have been used in prior studies to assess information quality across different media, including websites, social media posts, and videos (see section 2.2 for further details), also within the domain of nutrition and diet. We carefully examined each questionnaire, and grouped principles addressing the same or similar objectives and compared them across instruments. This process makes it possible to assess the completeness of each questionnaire and to identify the types of principles that best align with the characteristics of the analyzed data, which also includes video transcripts.

Among the reviewed tools, the Principles for Health-related Information on Social Media (PRHISM) questionnaire is found to best capture the aspects of information quality that are most measurable in the context of YouTube videos. In addition, it is well suited to social media platforms, as it was specifically designed for this type of content.

Several of the other examined questionnaires include principles that are primarily intended for other content formats, mainly written materials, and are therefore difficult to

adapt or measure using the available data. Examples include *Written using short sentences* and *Contain a space to make notes* from the EQIP questionnaire; *Mechanism to search, consult and locate the contents of the website* from Guardiola-Wanden-Berghe et al. [2011]; and *Uses peer review or another form of content review to vet information before sharing* and *Links to and is linked to by other credible sources* from the HONcode.

The questionnaires most closely aligned with PRHISM are DISCERN and the Social Media Evaluation Checklist. We excluded DISCERN because it is an older instrument, characterized by a lower level of detail and limited suitability for social media content. We instead excluded the Social Media Evaluation Checklist due to its weaker overall completeness.

Therefore, we selected the PRHISM questionnaire proposed by Denniss et al. [2022] as the primary assessment instrument and further complemented it with additional principles considered relevant for the purposes of this study.

Questionnaire integration and adaptation process

Several relevant and adaptable principles not included in PRHISM, but present in other questionnaires, were identified:

1. The aims are clear (DISCERN, EQIP);
2. The content achieves its stated aims (DISCERN, EQIP);
3. Acknowledges the limitations and evolution of knowledge (HONcode);
4. Keeps health information separate from financial, political, or ideological messages (HONcode);
5. Provides evidence to justify claims (HONcode);
6. Does not promote stereotypical attitudes or discriminatory actions related to clinical or social issues (Social Media Evaluation Checklist).

To improve completeness and ensure alignment with the study objectives, we aggregated original principles 3, 4, 5, and 6 into three additional principles and integrated them into the PRHISM questionnaire, resulting in a total of 16 principles used to assess information quality.

We did not include 1 and 2, as such principles were explicitly excluded during the development of the PRHISM questionnaire. According to the original study, these aspects were considered insufficiently important by the participants involved in the questionnaire design process (Denniss et al. [2022]).

Subsequently, we rephrased each of the 16 selected principles to adapt them to the context of diet- and weight-loss–related content on YouTube, while carefully preserving the original meaning. The complete list of rephrased principles is presented below.

Finalized principles

1. Authorship: It is clear whether the creator is speaking as a professional or as an individual without formal training and she/he clearly presents her/his identity and

relevant qualifications or experience (e.g., nutritionist, personal trainer, medical professional, or personal experience).

2. Authoritative: Information is within the speaker's scope of expertise. Professionals should speak only within their recognized field, while speakers who are not qualified or whose qualifications are unclear should not make technical, professional, or prescriptive claims.

3. Action-oriented: Information is presented clearly and concisely, providing enough context to help viewers make informed decisions, while acknowledging that multiple approaches may exist and explaining possible alternatives.

4. Financial disclosure: Clear and prominent disclosures of relevant sponsorship, advertising or financial support is disclosed. Paid sponsorships are made in a way that complies with the social media platform's guidelines.

5. Attribution: Information includes citations and hyperlinks (either verbally, in graphics, or in the video description) to the original source of information and includes the year the information was produced and/or published.

6. Balance and justifiability: Information is balanced and unbiased, avoiding exaggeration of benefits, minimization of risks, or promotion of 'miracle' solutions. The speaker does not promote restricted and/or excessive eating. Any claims are supported by evidence or clearly presented as personal experience.

7. Risks and benefits: Information clearly outlines the potential benefits and any associated risks or side effects of the products, behaviors, or practices being discussed.

8. Privacy: If information about a patient, client or third party is shared, it is shared with permission and does not include any identifying information.

9. Complementary information: Information is complimentary and not designed to replace the relationship between individuals and health professionals. Information includes statements encouraging individuals to discuss choices with a relevant health professional. Content that is primarily experiential and avoids prescriptive claims may still partially or fully satisfy this criterion, even in the absence of explicit disclaimers.

10. Referrals and support: The speaker directs viewers to additional reliable resources, official guidelines, or support services when relevant (e.g., eating disorder help lines).

11. Readability and comprehensibility: Information is presented in plain, everyday language, avoiding jargon, technical terms, abbreviations, or uncommon words, or providing clear explanations when such terms are necessary.

12. Accessibility: Information is accessible to vision and hearing-impaired individuals. Video contains closed captions and images include descriptive alternative text. Information is accessible with screen readers.

13. Images: Images and video clip accurately reflect the information presented, avoid being misleading, and are consistent with the verbal content of the video.

14. Acknowledgment of uncertainty: Video openly acknowledges the limitations of the evidence it presents and the fact that knowledge may evolve over time (e.g., preliminary or incomplete findings, small sample sizes, observational data that cannot establish causation, or rapidly changing topics).

15. Separation of interests: Information is separate from financial, political, or ideological messages, including avoiding promotion of stereotypical attitudes and discriminatory actions about clinical or social issues.

16. Data: Creator shares data to support statements.

Scoring process

To assess the information quality of each video, every principle included in the final questionnaire is scored on a scale from 0 to 4. If a principle is not applicable to the video content, it is marked as NA.

Each score corresponds to a specific level of criterion fulfillment, defined as follows:

0 = Completely unmet
1 = Very weakly met
2 = Weakly met
3 = Mostly met
4 = Completely met

As proposed in the study from Denniss et al. [2022], an overall information quality score is computed by averaging the non-NA scores across the 16 principles and subsequently rescaling the result to a 0–100 range. This procedure assigns each video a percentage value representing its level of information quality, referred to as the *quality score*.

3.6.2 Body-related questionnaire

We did not find any questionnaire to measure the level of body-related content in the literature. However, we used the *body-related variables* defined by Munro et al. [2024] as questionnaire. We kept each of the names of the 7 variables as criterion name and matched it with a short description adapt to the context and objectives and to the examples provided by Munro et al. [2024] themselves, to stick as possible to the idea behind the author variable creation. The principles finalized with this analysis are listed below.

Finalized principles

1. Weight measurement: Content includes references to body-weight measurement, whether as absolute values or as mentions of weight lost or gained. (ex. 'What I eat to lose 16 kg'.)

2. Mention of calories: Content includes mentions of the number of calories. (ex. 'This is one pint of ice cream, 1000 calories. And this is 8 cups of protein ice cream, 385 calories'.)

3. Referencing body image (ex. Man filming his shirtless body in the mirror at different timepoints. 'This is my body one day of dieting. One week of dieting. Two weeks of dieting'.)

4. Negative portrayal of body image: Content negatively refers to or negatively portrays body image. Body image is how we think and feel about ourselves physically (including our perceived sexual attractiveness) and how we believe others see us. (ex. A man starts working out to look more like the muscular men on TV his partner was admiring.)

5. Positive portrayal of body image: Content positively refers to or positively portrays body image. Body image is how we think and feel about ourselves physically (including our perceived sexual attractiveness) and how we believe others see us. (ex. 'This outfit looks so good on me, I don't care what you say, this dress looks so good on me')

6. Body checking: Creator repeatedly checks her/his shape or weight in the mirror or using the phone. (ex. Girl posing in front of mirror with stomach exposed and filming herself.)

7. Comparison: Content includes comparisons of the body across at least two time points, eventually using body-checking techniques to illustrate changes in weight or appearance, for example, Day 1 of a diet plan vs. Day 20. (ex. 'Diet journey. 83 kg59 · 6 kg'. Photos of a woman at different timepoints, with the weight at the time in text.)

In principles 4 and 5, we included the definition of *body image* provided by Munro et al. [2024], to guide labelers and LLMs (see following sections) in its understanding and improve accuracy of the scores.

Scoring process

We applied the same procedure used for the assessment of information quality to compute an overall body-related content score for each video. Specifically, each principle is assigned a score ranging from 0 to 4, or *NA* when not applicable, and the average score is subsequently rescaled to a 0–100 range, following the approach suggested by Denniss et al. [2022] and described in section 3.6.1. The resulting metric is referred to as the *body score*.

In this context, the levels of criterion fulfillment are defined as follows:

0 = None
1 = Rare, implicit, or fleeting
2 = Occasional, subtle, not central

3 = Frequent and noticeable
4 = Consistent or central throughout the video

3.7 Labeling scheme and manual annotation

After defining the principles of the two questionnaires, together with the corresponding scoring procedures and scales, the next step consists in assigning these scores to the videos included in the dataset.

To this end, we developed a structured labeling scheme. It includes the 23 principles derived from the two questionnaires, as well as two questions aimed at validating the NMF-derived topics, two questions designed to characterize the channel associated with each video, one question assessing the extent of AI use in the video production, and three questions recording potential brand or product mentions within the video. We further complemented the labeling scheme with the list of the 13 NMF topic names.

The two channel-related questions are multiple-choice items: the first concerns the type of channel owner, while the second addresses the channel category. The lists of response options are informed by the studies of Mejova and Tizzani [2025] and Gkouskou et al. [2011]. In a second stage, we introduced the channel category *Lifestyle* following an open-coding approach (Charmaz), based on insights gained during the exploratory phase and the initial labeling of videos conducted by the author.

Due to limited resources in terms of human annotators, it is not feasible to manually label all 3,129 videos. Therefore, we extracted a sample of 50 videos, with 25 randomly selected from the dynamic collection and 25 from the retrospective collection. We performed an additional subsampling, selecting 10 videos from each of the two 25-video sets, in order to obtain two groups that were as heterogeneous as possible in terms of NMF-assigned topics and video durations.

The author manually labeled all 50 sampled videos according to the defined labeling scheme. In addition, four external annotators independently labeled the subset of 20 videos selected in the second sampling step, dividing them into groups ranging from two to nine videos each. At the time the additional annotators began their work, the *Lifestyle* category had already been incorporated into the labeling scheme.

As a result, 30 out of the 50 videos were labeled only once, while the remaining 20 received double annotations, enabling the computation of inter-rater agreement, through Spearman's rank correlation coefficient (A.1). For these 20 videos, the responses to each question and principle included in the labeling scheme were also discussed between the author and the respective annotators in order to reach a consensus on every item. With the subsequent involvement of an LLM for video annotation in mind (3.8), during these discussions we decided to remove 5 of the 23 questionnaire principles. In particular, *Accessibility* (12) and *Images* (13) from the quality questionnaire, as well as *Body checking* (6) from the body-related questionnaire, were excluded because they relied heavily on visual information from the videos, which was not available to the LLM. Since the evaluation of body references typically relies more heavily on visual content, the outcomes derived from the questionnaire should be interpreted as underestimates. Additionally, we removed

Financial disclosure (4) and *Privacy* (8) from the quality questionnaire, as the available information was deemed insufficient to assign reliable scores, even for human annotators. The agreed-upon single-principle scores for the quality and body-related questionnaires, after the removal of these principles, are hereafter referred to as *agreement_scores*.

Based on the insights that emerged during these discussions, the author subsequently reviewed and, where necessary, revised the previously assigned scores for the 30 single-labeled videos to ensure greater consistency in the overall labeling process. These revised scores, on the non-deleted principles, are hereafter referred to as *author_scores*.

The additional questions included in the labeling scheme, beyond those derived from the quality and body-related questionnaires, are reported below.

Additional labeling questions

Content type

- Considering the full list of available topics, how accurate is the algorithm's selection of topics for this video and how appropriate are the assigned intensity levels in representing their relative relevance? (Give a score between 0 and 4)
- Please indicate the main topic or topics of the video using one word or a short expression (maximum 6 words in total).
  You may (a) use only labels corresponding to the listed topics if they fit the video content, (b) use only new labels if none of the listed topics adequately captures the video, or (c) combine listed topic labels with new labels. Any new labels should match the level of detail of the existing topics.

Channel type

- Which of the proposed options best describes the owner of the channel to which this video belongs?

  1. Institution (e.g. government, hospital or university)
  2. Commercial (e.g. sponsored site or private medical site)
  3. Individual (e.g. blogger)
  4. Other

- Which of the proposed options best describes the category of the channel to which this video belongs? (The definition of the category is extrapolated from the primary focus of the channel, as indicated by the thematic content of the videos produced and the content creators' self-declaration of intent.)

  1. Science (selfdeclared with a focus on science)
  2. Doctor (hosted by self-identified doctors)
  3. Health (focusing on healthrelated topics)
  4. News (self-declared news outlets)
  5. Opinion (not self-declared as official news or information sources)
  6. Lifestyle (self-referential content centered on lived experiences and daily practices)

7. Other

AI use

- Artificial intelligence tools have been used to generate or modify the video (including visual, textual, or audio elements such as AI-generated voice or sound).
  Give a score between 0 and 4, where

  0 = No AI tools were used in the video's production.

  1 = AI tools were used minimally, with negligible influence on the video's content (e.g., minor edits or isolated elements).

  2 = AI tools were used to modify or to generate limited portions of the video, without substantially shaping the overall content.

  3 = AI tools were used to modify or generate a substantial portion of the video, significantly influencing its content, though the video was not entirely AI-generated.

  4 = The video was entirely generated using AI tools.

  NA = AI use was difficult to detect

Product/Brand mentions

- Determine whether the content mentions at least one specific brand or branded product. Branded products include named commercial products such as (but not limited to) medications, apps, devices, or services. Generic or descriptive terms without a commercial name are not considered brands. The names of social media platforms or networks (e.g., Instagram, TikTok, YouTube) should not be counted as brands for this purpose.

- (If yes) Which products or brands are mentioned? (list them in lowercase, separated by commas. If the same entity is mentioned multiple times, list it only once.)

- (If yes) What exact words or phrases are used to refer to them? Report the exact string.

3.8 LLM-based automated labeling and validation

Manual labeling allowed the annotation of only 50 out of the 3,129 videos included in the dataset, which is insufficient to derive robust and generalizable insights from the subsequent analyses. To scale the annotation process, we employed large language models (LLMs) (Khalil et al. [2025]) in order to leverage their built-in knowledge, given the limited size of the labeled sample and the lack of resources required to train or fine-tune dedicated models.

In particular, we used OpenAI's GPT-4.1 model to assign 0–4 scores to the selected principles of both questionnaires for every video in the dataset, through `OpenAI` API.

To identify the most suitable model configuration and prompting strategy, we used 17 videos with double manual annotations, together with 20 additional videos randomly sampled from the 30 single-annotated ones, as a validation set (37 videos in total). We used the remaining 3 double-annotated videos as examples for the few-shot prompting procedure (3.8.1), taking into account their *agreement_scores*. We held out the remaining 10 manually labeled videos as an independent test set.

On the validation set, we applied different model configurations and prompting techniques. For each attempt, we computed *quality score* and *body score* as the rescaled average of the respective principle-level scores generated by the model. We then compared these overall scores with the corresponding human annotations, derived from either the *agreement_scores* (for double-labeled videos) or the *author_scores* (for single-labeled videos).

We conducted the comparison using Spearman's rank correlation coefficient, as preserving the relative ordering of videos between human and LLM assessments was considered more important than matching absolute score values. Further details on the correlation metric are provided in section A.1.

We compared the Spearman coefficients obtained from each LLM configuration with the coefficient resulting from the inter-rater agreement, in order to assess the extent to which the model could approximate human scoring behavior. The final selection of the model and prompting strategy was based on a joint evaluation of Spearman correlation, computational efficiency, and economic cost. We subsequently applied the selected configuration to annotate the entire dataset.

A detailed description of each LLM attempt on the validation set is provided in the following subsections.

3.8.1 Prompting strategies

In this study, prompting refers to the design of the textual instructions provided to the LLM in order to guide the scoring process. This includes the formulation of the task description, the specification of the scoring scale, and the inclusion of contextual information (i.e., title, description, and transcript). Variations in prompting strategy may influence the consistency, reliability, and reproducibility of the assigned scores and are therefore systematically evaluated.

A review of the literature on LLM-based annotation in comparable tasks (2.5) suggests that, alongside standard prompting, Chain-of-Thought (CoT) prompting in zero-, one-, or few-shot settings is among the most commonly adopted techniques (Wei et al. [2022], Kojima et al. [2022]).

Chain-of-Thought (CoT) prompting consists of explicitly encouraging the model to reason step by step before producing the final score. Rather than directly outputting a numerical value, the model is instructed to articulate intermediate considerations regarding the extent to which each principle is fulfilled, and only subsequently assign the corresponding score. The rationale behind this approach is that structured intermediate reasoning may enhance the alignment between model-based evaluations and human judgment. In this configuration, the GPT output does not consist solely of numerical scores, but also includes explanatory reasoning supporting each assigned value.

We also explored different shot-based paradigms. In zero-shot prompting, the model is asked to perform the task without being provided with any example of the expected input-output structure. In one-shot prompting, a single annotated example is included to illustrate the desired scoring behavior, whereas in few-shot prompting multiple annotated examples are provided. The inclusion of examples aims to guide the model toward a more consistent interpretation of the scoring criteria and output format. Unlike CoT prompting, one- and few-shot prompting do not necessarily require additional explanatory output compared to standard zero-shot prompting; however, they increase the number of input tokens due to the inclusion of example annotations.

Since the cost of GPT API usage is determined by the number of input and output tokens (with output tokens more expensive than input tokens) CoT prompting and shot-based approaches are economically more demanding than standard zero-shot prompting, despite their documented effectiveness.

To contain economical costs, we truncated transcripts longer than 4,000 words (approximately 4,000 tokens). Specifically, we retained only the concatenation of the first and last 1,500 words (3,000 words in total). We did not derive this threshold from experimental evidence or prior literature, but chose it pragmatically to maximize transcript coverage while remaining within the predefined budget constraints. For similar cost-related reasons, we did not implement CoT few-shot prompting; instead, a less expensive CoT one-shot configuration was adopted.

The prompting strategies that we ultimately evaluated are: zero-shot standard, few-shot standard, zero-shot CoT, and one-shot CoT prompting.

We performed all API calls using the `responses.parse()` function.

3.8.2 Prompt design and implementation

For each video, we performed three separate API calls: one dedicated to the body-related questionnaire principles and two addressing the quality questionnaire principles. We divided the quality principles into two groups based on thematic similarity, in order to cluster conceptually related criteria within the same call. The first group comprises principles 5, 16, 14, 10, 3, 11, and 9, whereas the second includes principles 1, 2, 15, 6, and 7.

The choice to process one video per API call is motivated by the need to preserve independence across samples as each API call is stateless and does not retain memory of previous requests. Ideally, each principle would be evaluated through a fully independent call; however, this approach is not economically feasible given the number of principles and videos. As a compromise, we divided the principles into three chunks, and the instruction "Evaluate each criterion independently" was explicitly appended to the prompt to mitigate potential cross-criterion interference within the same call.

Zero-shot prompting

An example of a zero-shot standard prompt is reported below:

You are given a YouTube video. Evaluate the content according to the criteria below.

Scoring scale (applies to all criteria):
-1 = Not applicable (the criterion does not apply to the video content)
0 = Completely unmet
1 = Very weakly met
2 = Weakly met
3 = Mostly met
4 = Completely met
Use only integer values from -1 to 4.

Criteria:
q14. Attribution: Information includes citations and hyperlinks ...
q25. Data: Creator shares data to support statements.
q23. Acknowledgment of uncertainty: Video openly acknowledges the limitations...
q19. Referrals and support: The speaker directs viewers to additional reliable resources...
q12. Action-oriented: Information is presented clearly and concisely, providing...
q20. Readability and comprehensibility: Information is presented in plain, everyday...
q18. Complementary information: Information is complimentary and not designed to...

Evaluate each criterion independently.
Base your evaluation on the title, description, and transcript.
Do not increase scores solely because the transcript is long.

Video id: <videoID1>
Title: The Fat Loss Mistake I Made for Years | 60 Pounds Down
Description: ♡ Video Links ♡ Snack Video. What I Eat in a Day...
Transcript: Hi guys, it's Mia. Welcome back to my channel. Welcome if you are new...

where principles and video details are partially cropped for presentation purposes, and the video ID is anonymized and replaced with a placeholder enclosed in angle brackets (<>).

The zero-shot CoT prompt follows the same structure, with the addition of the following sentence appended at the end of the instructions:

For each criterion, please explain your reasoning first (a couple of sentences) and then answer.

One- and few-shot prompting

We implemented one- and few-shot prompting through a manually managed conversation state using the `response` function. This allows additional messages to be provided as parameters of the function, thereby simulating a conversation with GPT that is taken into account during output generation.

We passed the task instructions (including task formulation, scoring scale, and output requirements) to the LLM as `developer` instructions. Subsequently, we constructed a simulated exchange between an imaginary `user` and the GPT `assistant`. In this simulated dialogue, the user provides information about one or more YouTube videos, and the assistant responds with the corresponding scores. These scores are, in reality, manually assigned human annotations and serve as reference examples to guide the model toward more accurate scoring of the final target video, for which human-assigned scores are not provided. The `developer` instructions are prioritized over the `user` messages within this setup.

An example of a simulated conversation for few-shot standard prompting is provided below:

developer: You are given a YouTube video. Evaluate the content...

user:
Video id: <videoID2>
Title: Growing Trend: Weight Loss Drugs and Type 1 Diabetes
Description: A new study finds the use of weight loss drugs like Ozempic...
Transcript: A new study finds the use of weight loss drugs like Ozempic,...
assistant: id: <videoID2>, q14: 0, q25: 0, q23: 2, q19: 1, q12: 3, q20: 3, q18: 4

user:
Video id: <videoID3>
Title: The Coffee Loophole for Weight Loss | Full Review & Recipe
Description: Learn the coffee loophole recipe here. UPDATE: The video will be...
Transcript: drinking coffee every day didn't work for me but it wasn't until I went...
assistant: id: <videoID3>, q14: 1, 25: 0, 23: 0, 19: 0, 12: 0, 20: 4, 18: 1

user:
Video id: <videoID1>
Title: The Fat Loss Mistake I Made for Years | 60 Pounds Down
Description: ♡ Video Links ♡ Snack Video. What I Eat in a Day...
Transcript: Hi guys, it's Mia. Welcome back to my channel. Welcome if you are new...

For representational purposes, only two examples are shown, whereas we used three shots in the few-shot standard configuration adopted in this study. Additionally, in the example above, the assistant's scores are presented on the same line, while in the actual implementation they are provided on separate lines.

Finally, an example of one-shot CoT prompting is reported below.

developer: You are given a YouTube video. Evaluate the content...
...For each criterion, please explain your reasoning first (a couple of sentences) and then answer.

user:
Video id: <videoID2>
Title: Growing Trend: Weight Loss Drugs and Type 1 Diabetes
Description: A new study finds the use of weight loss drugs like Ozempic...
Transcript: A new study finds the use of weight loss drugs like Ozempic,...
assistant:
id: <videoID2>
Reasoning14: Although the information is presented by a certified doctor, no sources...
14: 0
Reasoning25: Although the information is presented by a certified doctor, no data...
25: 0
Reasoning23: The speaker refers to the fact that doctors have changed their dosages...
23: 2
...
Reasoning20: The content is conveyed using simple language, avoiding abbreviations...
20: 3

user:
Video id: <videoID1>
Title: The Fat Loss Mistake I Made for Years | 60 Pounds Downr
Description: ♡ Video Links ♡ Snack Video. What I Eat in a Day...
Transcript: Hi guys, it's Mia. Welcome back to my channel. Welcome if you are new...

In this scenario, the reasoning provided for each principle score in the example was formulated by the author and the second video annotator. They reflect, as accurately as possible, the outcome of the discussion between the two evaluators that led to agreement on the assigned score for each principle.

3.8.3 Structured LLM outputs

In addition to specifying the task and scoring criteria within the prompt, the output format can be described textually in the prompt itself. A more robust solution is implemented through the use of a structured schema in Python.

Specifically, we employed the `BaseModel` class from the `Pydantic` library to define the expected data structure of the model responses. `Pydantic` enables automatic validation of data types and constraints, ensuring that the returned outputs conform to the predefined schema.

We defined separate child classes for each type of API call (first set of quality principles, second set of quality principles, and body principles), and for both standard and Chain-of-Thought (CoT) prompting configurations. In the standard setting, the schema allows only the return of the video identifier and the individual principle scores in integer format. In contrast, the CoT schemas additionally include a string field for the reasoning associated with each principle, thereby accommodating the explanatory component generated by the model.

Two examples of the defined output structures are reported below.

```
class ScoresQuality1(BaseModel):
id: str
q14: int
q25: int
q23: int
q19: int
q12: int
q20: int
q18: int
```

```
class ScoresQuality1Reas(BaseModel):
id: str
Reasoning14: str
q14: int
Reasoning25: str
q25: int
Reasoning23: str
q23: int
Reasoning19: str
q19: int
Reasoning12: str
q12: int
Reasoning20: str
q20: int
Reasoning18: str
q18: int
```

3.8.4 Model configuration

In addition to evaluating different prompting strategies, we tested two temperature settings of the model on the validation set. Temperature is a 0-2 parameter that controls the randomness of GPT's output, with lower values producing more deterministic and consistent responses and higher values generating more diverse and creative ones.

We tested a temperature of 1 based on prior literature (Khalil et al. [2025]), while a temperature of 0 is tried to favor reproducibility, given that creativity is not required for a structured scoring task. For each of the two temperature settings, we evaluated all four prompting strategies previously described (3.8.1).

Across all configurations, the adopted model is GPT-4.1. We specified the assistant's role through the `system` instruction: "Score the YouTube video based on criteria." We left all other model parameters at their default values.

3.9 Analytical framework

At this stage of the analyses, all the necessary information is available: video topics, non-semantic variables, and the quality and body-related content scores.

3.9.1 Q1: Topic modeling

We addressed the first research question (Q1) through the identification of the topics described in section 3.4; therefore, no further statistical analyses are required.

3.9.2 Q2: Determinants of quality and body-related content

We now turn to the second research question (Q2), which investigates the relationship between semantic and non-semantic video characteristics and the quality and body scores. To answer this question, we conducted univariate correlation analyses and estimated multiple linear regression models.

In this setting, we treated non-semantic variables (3.5) and video topics (3.4) as independent variables, while the previously described scores (3.6) as dependent variables. Our goal is to examine whether, and to what extent, the former are associated with the latter. We first estimated the univariate correlations and regression models using the *quality score* as the outcome variable, and subsequently using the *body score*. The analytical framework for Q2 is summarized in the following scheme:

- Video topics
- Non-semantic variables
- Channel characteristics
- Product/Brand mention
- AI-use

$\Big\}$ correlation & regression $\longrightarrow$ Information quality score, Body-related content score

where *Channel characteristics*, *Product/Brand mention*, and *AI-use* denote the categorical labels assigned by the human annotators on the manually evaluated subset, according to the adopted labeling scheme, in order to further enrich the characterization of the videos. When relationships involving these variables are analyzed, only the 50-video subset is considered; however, we still use the LLM-assigned scores to ensure comparability with the other analyses. Further details on these variables are provided in section 3.7.

Univariate analyses

Concerning univariate analyses between numerical independent variables and the outcome, we employed Spearman's rank correlation coefficient, as described in section A.1. We assessed statistical significance of the estimated coefficients through hypothesis testing with Bonferroni correction for multiple comparisons (section A.3).

When categorical independent variables are considered, particularly `channelCategories`, `channelCountry`, `categoryName`, human-labeled channel characteristics, and product/brand mention, Spearman's rank correlation cannot be computed. In these cases, we adopted the Mann–Whitney U test and the Kruskal–Wallis H test to evaluate whether videos belonging to different categories exhibit different distributions of *quality score* or *body score*, following the theoretical framework outlined in section A.3.

Since `channelCategories` and `channelCountry` exhibit a large number of distinct classes, we grouped categories with a video count below a predefined threshold into a single category labeled *Other*, which also includes videos with missing values. For `channelCategories`, we set the threshold to 40 videos, resulting in the top 17 categories being retained as separate classes. For `channelCountry`, we set the threshold to 10 videos, preserving the 14 most frequent countries as distinct categories.

We applied the Kruskal–Wallis test only when analyzing `channelCountry` and `categoryName`, as for the remaining categorical variable videos may be assigned to multiple

classes, violating the assumption of independence between groups. Since the test indicated statistical significance for both variables, we conducted post-hoc comparisons using the U test.

We performed the Mann–Whitney U test by considering one category at a time and comparing the group of videos belonging to that category with the complementary group of videos not belonging to it. We adopted this approach because the objective is to assess whether membership in a given category is associated with differences in the outcome distribution, rather than to conduct pairwise comparisons between all possible category combinations.

We applied the same Mann–Whitney procedure for `channelCategories` and for all tests performed, and evaluated statistical significance using Bonferroni-adjusted p-values.

We considered topic variables both as numerical and as categorical. In the numerical specification, each topic is treated as a separate variable, with values corresponding to the weights assigned by the NMF model to that topic for each video in the dataset. In the categorical specification, each video is associated with zero, one, or multiple predominant topics (see subsection 3.4.1). Accordingly, we performed both Spearman's rank correlation analyses and Mann–Whitney U tests in this setting.

We also performed Kruskal–Wallis and Mann–Whitney tests on categorical variables manually labeled in the smaller 50-video sample, specifically channel category, channel owner type, and the binary presence of product or brand mentions, following the same procedure and purpose described above. In contrast, we treated the extent of AI use as a numerical variable, and we assessed its relationships with other variables using Spearman's correlation.

Multiple linear regression

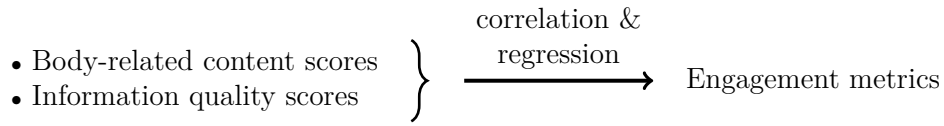
Subsequently, we employed multiple linear regression to assess the relationships when all independent variables are considered simultaneously. In this framework, topic variables are treated as numerical predictors, while we encoded purely categorical variables using dummy variables. For `channelCountry` and `categoryName` we used the *Other* category as the reference category, while, for `categoryName` we arbitrarily selected the category *Nonprofits & Activism* as the reference category. As `channelCategories` is concerned, as a single video may be associated with multiple channel categories, we adopted a multi-label dummy encoding.

Prior to model estimation, we applied an iterative variable selection procedure based on the Variance Inflation Factor (VIF) to reduce multicollinearity among the predictors. In addition, we standardized the input variables using a standard scaler in order to place them on a comparable scale, facilitating the interpretation and comparison of regression coefficients. After fitting the model, we conducted statistical tests to evaluate the significance of the regression coefficients. The theoretical framework underlying the linear regression procedure is detailed in section A.2.

3.9.3 Q3: Quality, body-related content, and user engagement

To address the third research question (Q3), which focuses on user engagement, we performed analogous univariate tests and regression analyses.

In this case, the quality and body-related content scores act as independent variables, whereas engagement metrics are considered dependent variables, as illustrated in the following schema:



In this context, we examined the relationships both using the aggregated *quality score* and *body score*, and by considering the individual scores of the single questionnaire principles.

The two engagement metrics that we considered in this analysis are the view count measured 70 days after upload and the engagement rate, defined as:

$$\text{engagementRate} = \frac{\text{likeCount} + \text{commentCount}}{\text{viewCount}},$$

where all counts are recorded 70 days after upload. The first metric captures engagement in absolute terms, whereas the second provides a relative measure of engagement.

We conducted additional correlation analyses to examine the relationships between video topics, numerical non-semantic features, and engagement metrics. The aim of these analyses is to identify potential engagement patterns associated with specific topics and metadata, and to evaluate whether these variables may act as mediating factors in the relationship between the *quality score* and *body score* and the observed engagement metrics.

3.9.4 Q4: Predicting risk-related scores

The final research question (Q4) explores whether it is possible to accurately predict information quality and level of body-related content through standard and cheap models, based on semantic and non-semantic variables that are easier to compute, thereby providing insight into the potential presence of harmful content.

To this end, we employed neural network models, specifically multy-layer perceptrons (A.4), as the task is inherently predictive and such models are well suited to capturing complex and non-linear relationships between input features and target variables. The analytical structure is therefore similar to that adopted for Q2, but framed as a prediction task.

Another difference with respect to Q2 is that we considered the dependent variables both in their continuous form, leading to regression models, and in a binary form, leading to classification models. We then compared the performances of these approaches. To obtain the binary formulation, we defined thresholds for both the *quality score* and the *body score*. These thresholds are intended to simulate potential decision boundaries that platform moderators might adopt to identify videos requiring further review. In this experimental setting, we arbitrary set thresholds at 20 for the *quality score*, where videos

with scores below 20 are considered low-quality, and at 80 for the *body score*, where videos with scores above this value are considered highly body-centered. It is important to note that these thresholds are chosen only for experimental purposes. In practical applications, they should be determined more carefully, taking into account the specific context of the platform and insights derived from past moderation practices.

In this setting, *Channel characteristics*, *Product/Brand mention*, and *AI-use* are excluded, since they are available only for a limited subset of videos, which would be insufficient for training a neural network whose performance strongly depends on the size of the training dataset. The corresponding schema is reported below.

- Non-semantic variables
- Video topics

$\Big\}$ NN (prediction) $\longrightarrow$ Body-related content score, Information quality score

In this configuration, we considered the categorical independent variables in their dummy representation, with missing values in channel categories filled with zeros. We instead filled missing values for numerical variables with the median of the non-null values, due to the high skewness of their distributions.

We tested different model configurations through a grid-search approach on a selected subset of hyperparameters. For each of them, we considered the default value provided by `scikit-learn` (the library used in this work), together with other close values. The hyperparameters and corresponding values that we considered for the grid-search are listed below, while we kept all the others at their default values.

- Hidden layer architecture (`hidden_layer_sizes`): combinations of number of layers (1, 5, 8, 10) and neurons per layer (20, 50, 100, 200)
- Activation function (`activation`): tanh, relu
- Strength of the L2 regularization term (`alpha`): 0.0001, 0.05
- Batch size (`batch_size`): 10, 50, auto

We separated from the whole dataset a test set with a dimension equal to 20% of the dataset length. We then fitted a standard scaler on the remaining 80% of the data (training set) and applied to transform those observations. We subsequently used the same learned scaler parameters to transform the values of the test set.

To select the best model, we then performed 5-fold cross validation on the samples not belonging to the test set and selected the model leading to the highest average recall score, for classification, or the highest average negative mean squared error for regression, on the validation portions of the cross validation.

We chose recall as the decision metric in the classification setting because it evaluates how many positive samples are correctly categorized as positive, giving less importance to negative samples that are erroneously classified as positive. In our scenario, in which the predictive model can be used to make a first selection of videos potentially risky that will subsequently pass through human review, it is more important to detect all risky videos than to avoid reviewing extra videos erroneously classified as risky.

We then re-trained the best selected model on the full 80% training sample and evaluated its performance on the held-out test set.

We repeated all the steps described above, from the grid-search definition to the evaluation on the test set, both with *quality score* and *body score* as dependent variables, and both in the classification and regression settings. Furthermore, for all the analyses described, we fixed the random seed to 42 to ensure reproducibility.

In the regression models, where the non-binarized *quality score* and *body score* are predicted, we subsequently applied a binarization to the predicted scores and computed classification performance metrics on these binarized values, in order to allow comparability with the classification models.

Chapter 4

Results

In this chapter, we report the results obtained through the methodology described in the previous chapter, starting with descriptive statistics of the collected and organized dataset. We then present the results of the topic modeling algorithm (section 4.2), addressing the first research question (Q1).

Section 4.3 reports the outcomes of both human and LLM-based labeling, showing the distributions of the assigned labels and presenting agreement metrics used to select the most suitable LLM model and to evaluate its ability to approximate human annotation behavior. This section lays the groundwork for the subsequent analyses, whose results are presented in section 4.4. Here, correlation and regression coefficients are reported together with the statistics of the tests performed to assess the relationships between metadata and the *quality score* and *body score*, allowing us to draw conclusions relevant to the second research question (Q2).

We addressed the third research question (Q3) in section 4.5, where we examined correlation and regression coefficients in relation to the *quality score*, *body score*, numerical metadata, and user engagement metrics. Finally, we report the selected hyperparameters and the performance of the predictive models that we developed to address Q4 (section 4.6).

4.1 Descriptive statistics

Through the dynamic collection described in section 3.1, we retained the initial metadata of 2,870 videos, uploaded by 1,681 distinct channels. Among these videos:

- 2077 have associated tags,
- 2602 have not-null descriptions,
- 2858 have not-null `topicDetails.topicCategories` fields,
- and 1,252 have a language specified in `snippet.defaultLanguage`.

Of the 2,870 videos, 1,157 were retrieved with the required frequency, meaning that all periodic metadata are available for them. These videos were uploaded by 798 distinct

channels. Among them:

- 824 have associated tags,
- 1034 have not-null description,
- 1153 have not-null `topicDetails.topicCategories` fields,
- and 522 have a language specified in `snippet.defaultLanguage`.

Some of the 1,157 videos may still be missing channel data or comment data for certain periodic checks. However, this was not considered relevant, since the project does not aim to analyse the temporal development of these variables.

Regarding the retrospective collection, 384 videos uploaded by 298 channels were retained. Among them:

- 261 have associated tags,
- 341 have non-null descriptions,
- and 382 have non-null `topicDetails.topicCategories` fields.

Overall, combining the dynamic and retrospective collections, results in a total of 3,129 videos from 1861 distinct channels.

4.1.1 Video duration and engagement

In the first analyses, video duration, engagement level and category are taken into consideration.

Table 4.1, concerning exclusively the data from the dynamic collection, shows that both video duration and, even more markedly, engagement metrics exhibit highly right-skewed distributions. This indicates that most videos are short and receive limited audience interaction, while a small subset accumulates disproportionately large values. Only 25% of the videos exceed 15 minutes in length, although durations can reach up to one hour.

Table 4.1: Summary statistics of video attributes, including duration, view count, like count, and comment count, for the *initial data* videos measured 70 days after upload. Duration is reported in seconds. Data after 70 days are computed based on 2658 videos over 2870 initial videos.

	duration	**view 70**	**like 70**	**comm. 70**
mean	656.48	16365.63	598.26	46.36
std	732.79	166780.22	5087.36	255.18
min	61.00	0	0	0
25%	154.00	76.25	3.00	0.00
50%	337.00	588.00	19.00	3.00
75%	900.25	4048.50	223.00	23.00
max	3593.00	5176342	144163	8219

The skewness is even stronger for engagement: the extremely long tail suggests that a small fraction of videos drives the majority of user activity. One day after upload, half

of the videos have 10 likes or fewer and 2 comments or fewer, with at least 25% having no comments at all. However, since all search queries were executed in the evening, these early engagement values are affected by the upload time of each video: videos posted early in the morning had nearly two full daylight periods to accumulate interactions, whereas those uploaded late at night had been public for less than 24 hours.

A more stable picture emerges when examining engagement after 70 days from upload, as these values are less sensitive to posting time and better reflect the settling dynamics of engagement metrics, whose growth rate decreases over time (as illustrated for view count in Figure 4.1). Even at this later stage, the distributions remain highly right-skewed. Half of the videos fall within a three-order-of-magnitude range of view counts (from 4,048.50 to 5,176,342), while 25% remain below 76.25 views. Comment activity is similarly sparse: at least 25% of videos still have zero comments, 50% have three or fewer, and only the top quartile exceeds 23 comments, with maxima reaching 8,219.

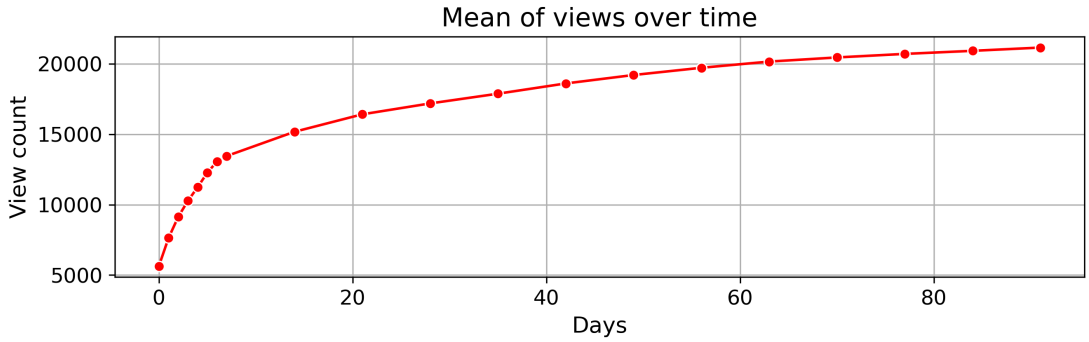


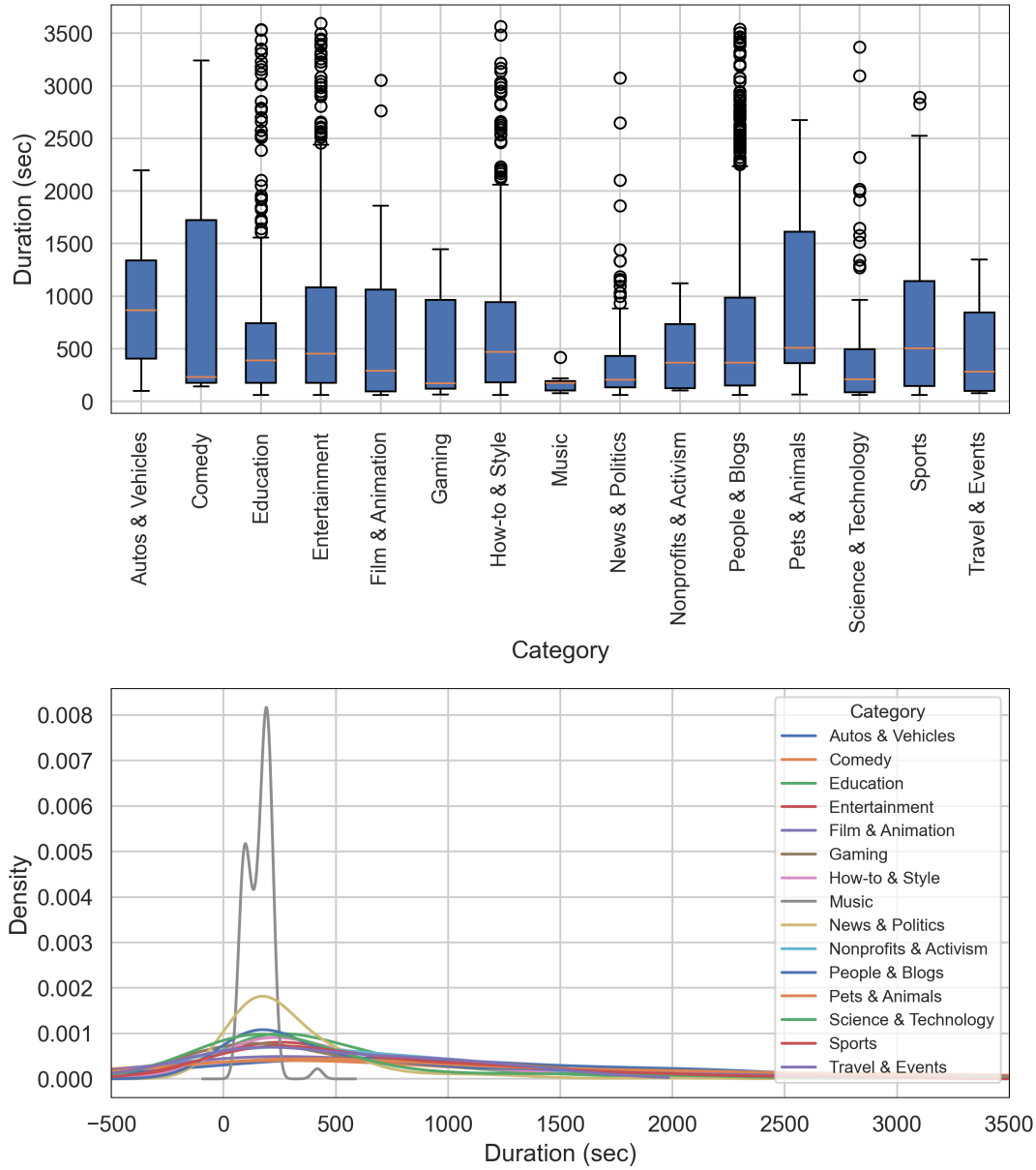
Figure 4.1: Mean view count across all *initial data* videos as a function of publication age.

These observations are consistent with the visual evidence provided by Figure 4.2, which instead concerns data from both collections. Although all categories exhibit a similar skewed shape, notable differences emerge. In particular, the *Music* category shows a concentration of relatively short videos, with a tight double-peaked distribution around lower duration values. Other categories tend to present slightly more spread-out distributions, but none reach the degree of dispersion observed in engagement metrics overall.

Categories such as *News & Politics* or *Science & Technology* generally show more moderate durations and a relatively narrow interquartile range, indicating that creators within these topics tend to produce videos of more uniform length. In contrast, categories like *Comedy* or *Pets & Animals* exhibit a much wider spread, with no strong central tendency. Notably, *Autos & Vehicles* and *Sports* include the highest mean video durations among all categories.

However, the varying number of videos assigned to each category, as reported in Table 4.6 and Figure 4.6, affects the reliability of these comparisons. Categories with low representation, such as *Nonprofits & Activism* (four videos), offer less statistically robust insights, and observations drawn from them should therefore be interpreted with caution. On the other hand, *People & Blogs*, the most represented category in the dataset, shows a particularly high number of outliers, reflecting the overall distribution of durations.

Figure 4.2: Box plot (top) and density estimates (bottom) illustrating the distribution of video duration across YouTube video categories in the *initial data*.



4.1.2 Channel data

The dataset shows a clear concentration of weight–loss and diet–related video content coming from a small set of countries, as illustrated in Table 4.2. The United States accounts for by far the largest share of channels (1024), followed at a considerable distance

by the United Kingdom (142) and India (107). This distribution is expected, given that the dataset was filtered by language and only English–language videos accessible from the United States were retained. As a consequence, English–speaking countries naturally occupy the top positions in the ranking, while the first country based in Central or South America, Mexico, appears only in ninth place with 12 channels. It can also be observed that some Central or South American countries, particularly Mexico and Brazil, exhibit a disproportionately high number of videos relative to the number of channels. This suggests that the discourse on the topic in these contexts is driven by a relatively small number of highly active channels.

Regarding thematic distribution (Table 4.3), the most represented channel categories are *Health*, *Lifestyle*, and *Physical fitness*. This suggests that weight–loss content is produced within a mix of lifestyle–oriented and health–oriented contexts, with the latter potentially including both medical guidance and more general wellness advice.
The prominence of categories such as *Physical fitness* and *Food* indicates that creators tend to emphasise practical routines and non-informational nutritional tips. Notably, *Knowledge* appears relatively low in the ranking, suggesting that explicitly educational or explanatory content plays a secondary role compared to routine-based or advice-driven formats. Conversely, categories such as *Politics*, *Film*, or *Pet* appear only marginally, which aligns with expectations for this type of content.

Table 4.2: Channel count (N_c) and video count (N_v) grouped by channel country. Shown 12 most frequent countries. 265 null-country channels.

Country	N_c	N_v
United States	1024	1636
United Kingdom	142	233
India	107	141
Canada	80	156
Australia	36	46
Pakistan	20	23
Nigeria	19	19
Germany	16	17
Mexico	12	85
Brazil	12	25
South Africa	9	9
Philippines	8	8

Table 4.3: Channel count (N_c) and video count (N_v) grouped by channel topic category. Shown 13 most frequent categories.

Topic	N_c	N_v
Health	1569	1943
Lifestyle (sociology)	1170	2400
Physical fitness	376	561
Food	297	585
Society	145	357
Entertainment	104	348
Television program	88	192
Knowledge	27	367
Pet	15	
Film	12	76
Hobby	10	58
Fashion	9	62
Politics	8	125

To compute statistics on the most popular and the most active channels, we selected the top 6% of channels based on subscriber count, views count, and number of uploaded videos. We then combined channels ranked by subscribers and views into a single category representing the most popular creators. Table 4.4 reports the channels that contributed the largest number of videos to the dataset, while Table 4.5 presents the same ranking restricted to the subset of the most popular channels.

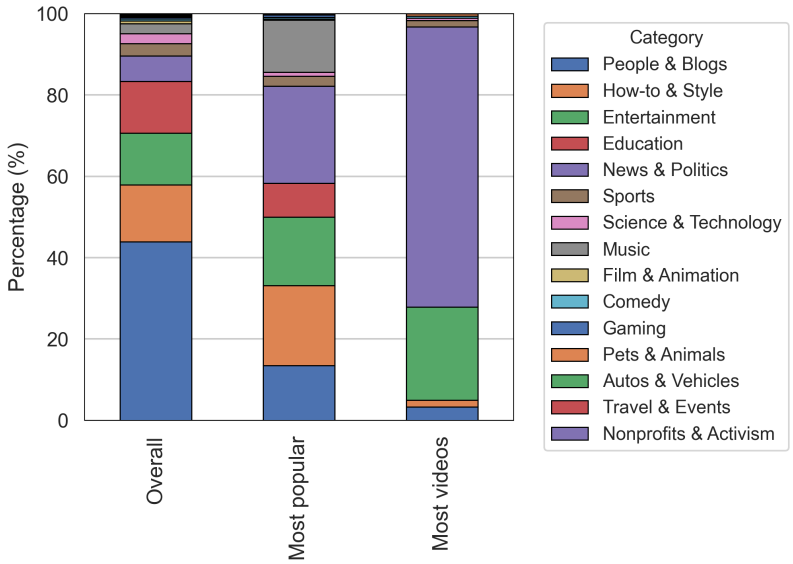
Table 4.4: Number of videos per channel title present in the dataset. Shown 13 most frequent channels.

Title	N
Wrestling Best	158
Janeé	47
Countess of Shopping	46
coldpizza.	37
Healthcare Tips and Aesthetics Advice	21
Infinite Letters	19
Kelly Story	19
Boreal	18
Ilmee Mintz	15
Sillz	15
Snake Diet	14
Lauren Jansen	14
Pris Tvv	14

Table 4.5: Number of videos per popular channel title present in the dataset. Shown first 13.

Title	N
coldpizza.	37
Infinite Letters	19
Versatile Vicky	13
E! News	10
Luis Mihajlow 1	9
v a l ; l e	7
CNBC Television	6
RookieSubs	5
Chef Ricardo Cooking	5
Yoga with Souvik	5
BrunoTraductor Official	5
This Morning	5
The Golden Balance	4

Figure 4.3: Category distribution of videos across all channels, the subset of most popular channels, and the channels with the highest number of uploaded videos. Percentages refer to the relative share of YouTube categories within each group.



The analysis of channel activity reveals notable differences between the overall dataset and the subset of the most popular creators. As shown in Table 4.4, the channels contributing the highest number of videos are not necessarily those with the largest audiences.

In contrast, the ranking of the most popular channels in Table 4.5 displays significantly lower video counts, indicating that high visibility and large subscriber bases do not correlate directly with the amount of weight–loss content produced. It is also worth noting that the two most prolific popular channels in terms of video count are both declared as being based in Mexico.

The differences are further reflected in the video category distribution displayed in Figure 4.3. The overall dataset exhibits a diverse mix of categories, with a strong presence of *People & Blogs* and *How-to & Style*, which aligns with the practical and routine-oriented nature of much weight–loss content. However, among the most popular channels, categories such as *Music* and *News & Politics* become relatively more prominent, with *News & Politics* being the most represented category over the channels with the highest number of uploaded videos.

Video creators in the *News & Politics* category, due to the typically shorter duration of their content, can afford a higher publication frequency, which likely contributes to the very high number of uploaded videos per channel in this category.

4.1.3 Other non-semantic variables

We present an overview of the distributions of additional non-semantic variables, mainly related to text lengths, in Figure 4.4. For a detailed description of these variables, see section 3.5.

The distribution of the number of characters in video titles is left-skewed, with a clear peak at exactly 100 characters, suggesting the presence of a platform-imposed limit on title length. In contrast, `description_len`, `transcript_len`, and `channelDescription_len` exhibit strongly right-skewed distributions, reflecting the predominance of shorter descriptions and transcripts, and aligning with the previously observed right-skewed distribution of video durations.

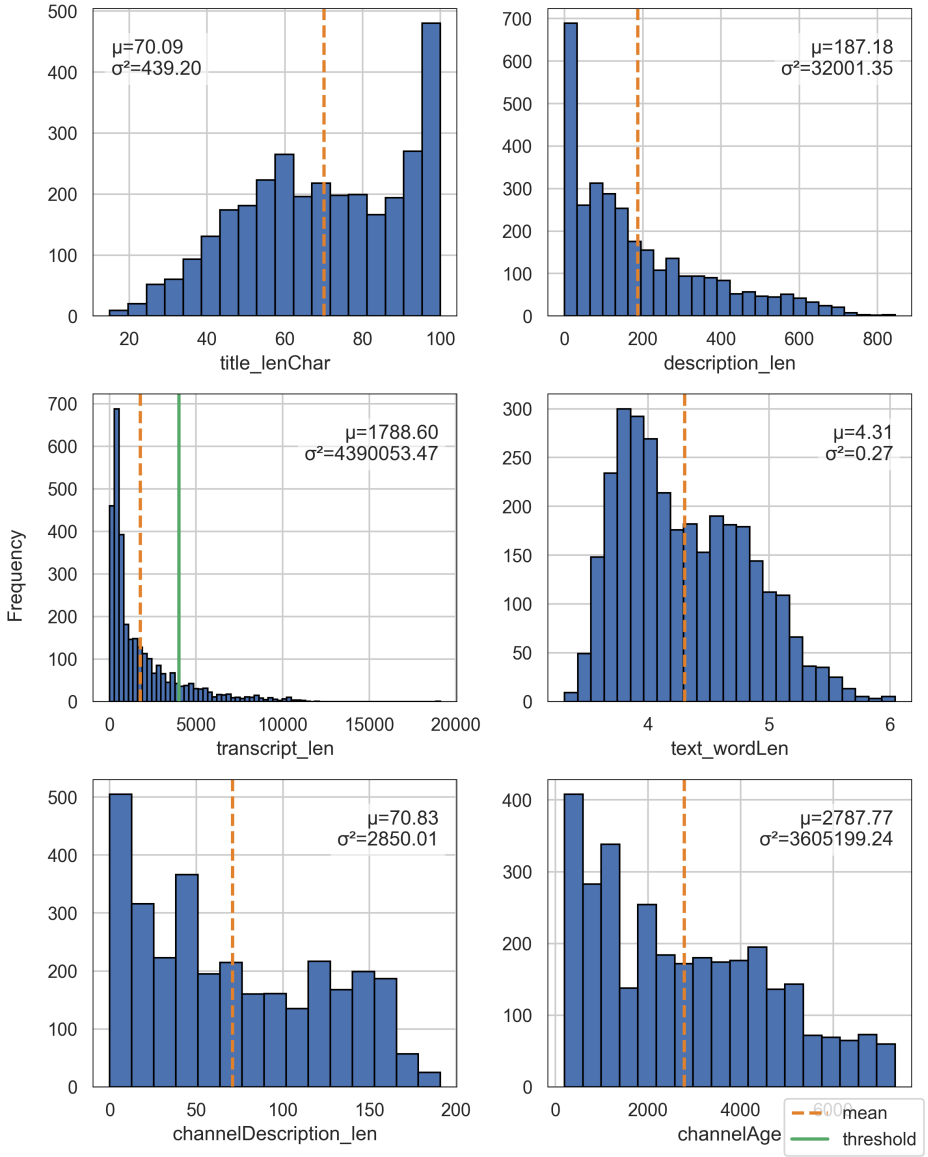
A similar, though less pronounced, skewness can be observed for channel age, with a relatively high number of channels created shortly before the publication of the analyzed video.

Finally, `text_wordLen` displays a distinctive bimodal distribution, with a more prominent peak around 3.8 characters and a secondary, smaller peak around 4.7 characters.

4.1.4 Correlations between non-semantic variables

Considering Spearman's correlations among the non-semantic numerical variables, the only notable negative coefficient (approximately -0.50) is observed between `text_wordLen` and `duration_secs`. This indicates that videos with longer average word lengths tend to be shorter in duration, and vice versa. This somewhat counterintuitive result may be influenced by the presence of the topic *Personal storytelling*, discussed later (see section 4.2). A large proportion of the longer videos in the dataset belong to this topic, suggesting that the use of shorter words may be more characteristic of this narrative style rather than a general feature of longer videos.

Figure 4.4: Distribution of six non-semantic variables across the videos in the dataset, with each plot showing the corresponding mean (μ) and variance (σ^2). The green vertical line marks the transcript-length threshold of 4,000 words, above which transcripts were truncated for LLM labeling (3.8.1).



Among the positive correlations, a coefficient of 0.65 is found between `channelSubscriberCount` and `channelVideoCount`, indicating that channels with more uploaded videos generally have a larger subscriber base. Additionally, a correlation of 0.55 is observed between `description_len` and `description_links`, implying that longer

descriptions are more likely to include a higher number of links. All other correlation coefficients between non-categorical variables remain below 0.5 in absolute value.

When categorical features are also taken into account, unexpected patterns emerge. Channels declared as based in Mexico show moderate positive correlations with music-related channel categories, including *Independent music* (0.55), *Music* (0.49), *Music of Latin America* (0.49), and *Pop music* (0.56). None of the other countries exhibit correlations greater than 0.25 in absolute value with any relevant channel category.

A further positive correlation of 0.49 is observed between Mexico and the video category *Music*. In the correlation matrix between channel countries and video categories, the next highest coefficients (in absolute value) are substantially lower: 0.17 between Peru and *Music*, and 0.15 between Brazil and *Music*. These findings suggest that, in countries based in Central or South America, particularly Mexico, channels publishing diet- and weight-loss-related videos are often primarily music channels. In contrast, in other countries, such content is only rarely associated with music channels.

Keyword	Video count
weight loss	2216
weight	1647
loss	703
lose weight	220
loss journey	182
day	174
yeah	164
people	146
mitolyn	124
week	120
body	117
fat	113
music	98
protein	88
eat	86
good	83
things	80
calories	76
health	74

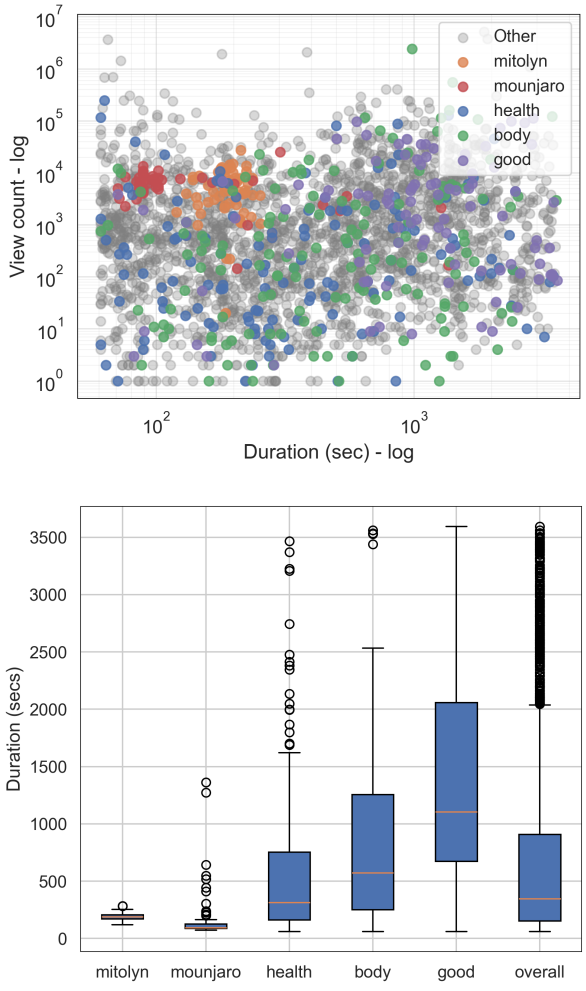


Figure 4.5: Keywords and corresponding video counts (top). Relationship between video duration and view count after 70 days across keywords (top right). Video durations per keyword (bottom right).

4.1.5 Keyword analysis: the Mitolyn and Mounjaro case

The ranking of most frequent keywords extracted through YAKE! (see section 3.3), is shown in Figure 4.5. Excluding trivial and uninformative keywords, and taking into consideration the research questions of the study, we selected five keywords: *mitolyn*, *mounjaro*, *health*, *body*, and *good*. Surprisingly, two out of five, are pharmaceuticals: Mitolyn is a dietary supplement, while Mounjaro a diabetes medicine. Even more surprising is the distribution of view counts after 70 days from upload and durations of videos associated to each of the five keywords or keywords including those terms, shown in the top right graph of Figure 4.5. Differently from videos related to *health*, *body*, and *good*, the ones associated to the two drugs present similar values of view count and duration, collocating themselves in two well-defined and compact regions of the scatter plot. The bottom-right graph in Figure 4.5 shows even more clearly how videos associated with these two keywords are much shorter and have similar durations to each other.

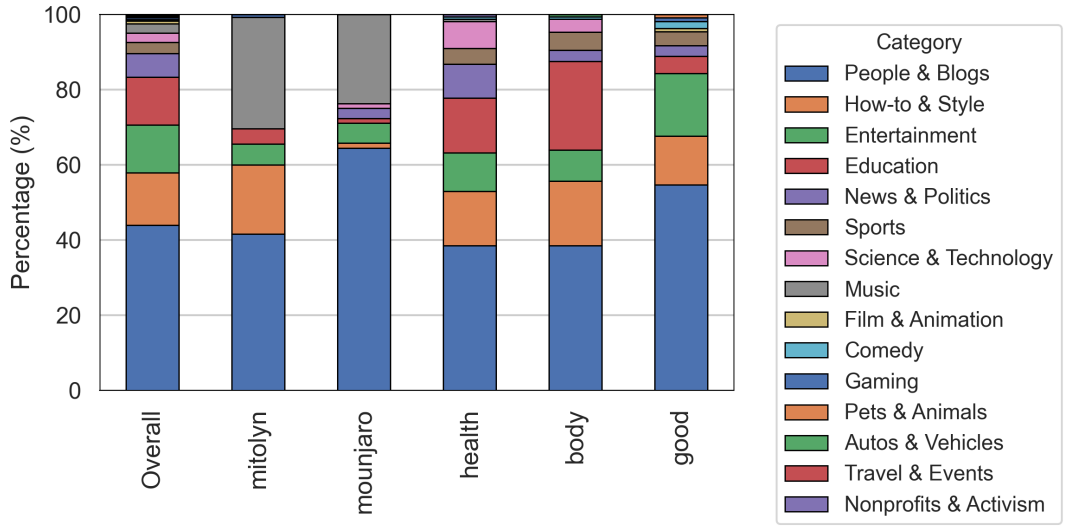


Figure 4.6: Category distribution of videos associated with each keyword. For each bar, the stacked segments represent the proportion of YouTube categories among the videos retrieved with the corresponding keyword (or keyword including the term).

What is noteworthy in Figure 4.6 is the increase of the *Music* category among the videos that mention the two drug-related keywords, *mitolyn* and *mounjaro*, accompanied by a decrease in the *Entertainment*, *Education*, and *News & Politics* categories.
Given this counterintuitive pattern, and the unexpected prominence of the *Music* category among diet- and weight-loss–related videos, a more detailed investigation was conducted. Specifically, explicit keyword occurrences were searched for in the concatenated text of each video (title, description, and transcript). The keywords examined were: *mitolyn*, *mounjaro*, *product*, *merch*, and *pill*. Table 4.6 reports the percentage of videos mentioning each keyword, broken down by category, where *prod* refers to the presence of at least one of the listed keywords.

A suspicious pattern emerges again for the *Music* category, which shows the highest proportion of videos containing product-related terms. It ranks first for mentions of *mitolyn*, reaching 50.65%, compared with only 7.78% in the second-highest category. *Music* also leads in mentions of *mounjaro* with 28.57%, followed by *People & Blogs* with 12.37%. The second category with the highest proportion of product-related keywords is *Autos & Vehicles*; however, it includes only 8 videos, none of which mention either *mitolyn* or *mounjaro*.

Table 4.6: Percentages of videos mentioning product-related keywords across categories. For each category, the table reports the total number of videos, the number and percentage of videos containing at least one product-related keyword (*prod*), and the percentages of videos explicitly mentioning *mitolyn* or *mounjaro*.

	N	N prod	% prod	% mitolyn	% mounjaro
Music	77	69	89.61	50.65	28.57
Autos & Vehicles	8	5	62.50	0.00	0.00
How-to & Style	437	208	47.60	7.78	4.58
People & Blogs	1374	629	45.78	4.22	12.37
Entertainment	400	161	40.25	1.75	8.50
Education	397	155	39.04	1.51	5.54
Gaming	13	5	38.46	7.69	7.69
Film & Animation	21	8	38.10	0.00	0.00
News & Politics	196	72	36.73	0.00	11.73
Science & Technology	78	27	34.62	0.00	10.26
Sports	94	28	29.79	0.00	0.00
Comedy	14	4	28.57	0.00	0.00
Travel & Events	7	2	28.57	0.00	0.00
Nonprofits & Activism	4	1	25.00	0.00	0.00
Pets & Animals	9	2	22.22	0.00	0.00
Overall	3129	1376	43.98	4.63	9.59

The two channels posting the highest number of videos associated with the keyword *mitolyn* (or keywords containing *mitolyn*) are *Wrestling Best*, with 34 videos, and *coldpizza.*, with 16 videos. The third and fourth channels in this ranking each have only 9 videos, highlighting the large gap separating the top two from the rest.
Among the 28 channels that uploaded videos associated with the *mitolyn* keyword (or related keywords), one channel's data could not be retrieved, and only one of the remaining channels was associated with the United States, despite US being the most popular country among the channels of the whole dataset (Table 4.2). Six had a missing country value, whereas the others were located in Central or South America, with Brazil (7 channels) and Mexico (6 channels) dominating the ranking. They are followed by the Dominican Republic, Peru, Chile, Ecuador, Venezuela, and Colombia.

The *coldpizza.* channel is no longer available, while we carried out a more in-depth manual analysis for *Wrestling Best* and for other top channels by number of *mitolyn*-related videos. This inspection revealed that several of these channels were originally

created to post music content (and some still appear to do so), but were later repurposed to share videos promoting pharmaceuticals. Notably, the same three women appear in multiple videos across different channels, suggesting the presence of a single producer behind this content.

A similar phenomenon is observed for the keyword *mounjaro*, which often appears in the expression *natural mounjaro*. The videos associated with this term typically feature recipes for preparing a drink claimed to have effects similar to the Mounjaro drug. Once again, the same women appear to be behind this content.

4.1.6 Keyword analysis: body-related content and disclaimer

We carried out a keyword-occurrence count, similar to the one that we used for detecting product mentions (subsection 4.1.5), using weight- and calorie-related keywords. Specifically, we searched for explicit mentions of the following terms for weight-related content: *kg*, *kilo*, *lb*, and *pound*. For calorie-related content, we chose the keywords *calorie* and *kcal*. The results are reported in Table 4.7.

Table 4.7: Percentages of videos mentioning body-related keywords across categories. For each category, the table reports the total number of videos, the number and percentage of videos containing at least one weight-related keyword (*weight*), and the number and percentage of videos containing at least one calorie-related keyword (*calories*).

	N	N weight	% weight	N calories	% calories
Autos & Vehicles	8	6	75.00	2	25.00
Comedy	14	7	50.00	7	50.00
Education	397	249	62.72	201	50.63
Entertainment	400	279	69.75	153	38.25
Film & Animation	21	16	76.19	5	23.81
Gaming	13	8	61.54	3	23.08
How-to & Style	437	296	67.74	226	51.72
Music	77	50	64.94	52	67.53
News & Politics	196	128	65.31	21	10.71
Nonprofits & Activism	4	3	75.00	0	0.00
People & Blogs	1374	979	71.25	696	50.66
Pets & Animals	9	6	66.67	4	44.44
Science & Technology	78	49	62.82	29	37.18
Sports	94	67	71.28	56	59.57
Travel & Events	7	4	57.14	3	42.86
Overall	3129	2147	68.62	1458	46.60

Across all categories, a substantial proportion of videos contains explicit references to weight-related terms, with an overall rate of 68.62%. Categories such as *Film & Animation* (76.19%), *Autos & Vehicles* (75%), *Nonprofit & Activism* (75%), and *Sports* (71.28%) show the highest percentages, although some of these categories include only a small number of videos and weight-related measurements may refer to objects or animals rather

than humans. More informative are the larger categories: *People & Blogs* (71.25%), *Entertainment* (69.75%), and *How-to & Style* (67.74%), all of which display consistently high frequencies of weight-related mentions, confirming their centrality in hosting weight- and diet-related content.

Mentions of calorie-related terms show more variability across categories, with an overall proportion of 46.60%. The category *Music* stands out with 67.53%, despite its relatively low relevance to weight loss content in principle, reinforcing earlier observations about unexpected patterns in this category. Other categories with substantial representation include *Sports* (59.57%), *How-to & Style* (51.72%), *People & Blogs* (50.66%), and *Education* (50.63%). By contrast, categories such as *Nonprofits & Activism* (0%) and *News & Politics* (10.71%) exhibit low frequencies of calorie-related keywords, reflecting their different thematic focus.

We also performed the keyword search on the concatenated text of each video using the term *disclaimer*, and the corresponding results are reported in Table 4.8.

Table 4.8: Number and percentage of videos mentioning the *disclaimer* keyword across categories.

	N	N discl.	% discl.
Travel & Events	7	2	28.57
Science & Technology	78	19	24.36
How-to & Style	437	94	21.51
Film & Animation	21	4	19.05
Education	397	73	18.39
Sports	94	17	18.09
People & Blogs	1374	224	16.30
News & Politics	196	25	12.76
Autos & Vehicles	8	1	12.50
Entertainment	400	47	11.75
Gaming	13	1	7.69
Comedy	14	0	0.00
Music	77	0	0.00
Nonprofits & Activism	4	0	0.00
Pets & Animals	9	0	0.00
Overall	3129	507	16.20

Across categories, the explicit mention of the *disclaimer* keyword appears in a relatively small proportion of videos, with an overall rate of 16.20%. The highest percentages are found in *Travel & Events* (28.57%), *Science & Technology* (24.36%), and *How-to & Style* (21.51%). Moderate values are observed in categories that also contain a substantial number of videos, such as *People & Blogs* (16.30%). Conversely, several categories, including *Music*, *Comedy*, *Nonprofits & Activism*, and *Pets & Animals*, contain no videos mentioning the keyword. These results indicate that the explicit mention of disclaimers is not widespread across the dataset and tends to appear slightly more often in instructional, informational, or professionally oriented content, while being largely absent in

entertainment-focused categories.

This information should be interpreted with caution, as the keyword may refer to disclaimers of any kind, not necessarily those related to diet or weight loss. For example, in the description of one of the videos that matched the keyword search, the following statement can be found:

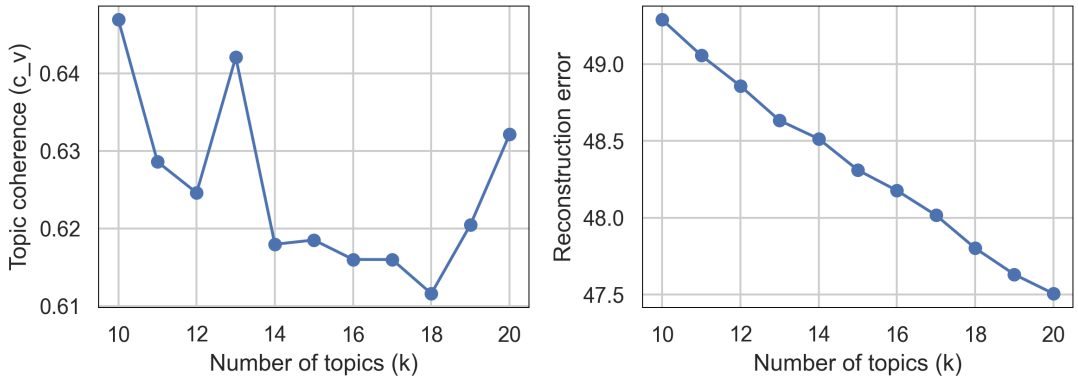
> *Copyright* ***Disclaimer*** *under section 107 of the Copyright Act 1976*
> *We do not own the rights to all content. They have, in accordance with fair use, been repurposed with the intent of critical analyst and review.*
> *We must state that in NO way, shape or form am I intending to infringe the rights of the copyright holder. Content used is strictly for research/reviewing purposes and to help educate. All under the Fair Use law.*

The data can still provide an indication of how much attention creators typically pay to disclaimers, which can still be interesting.

4.2 Q1: Topic modeling

To obtain an overview of the themes discussed in the collected videos, we performed topic modeling using Non-negative Matrix Factorization (NMF) (3.4). When applying NMF, we tested different values of max_df and k in order to identify the model achieving the highest coherence and best human interpretability, following a grid-search approach. The results of this search indicated $max\_df = 0.375$ and $k = 13$ as the optimal combination of hyperparameters.

Figure 4.7: Average topic coherence and reconstruction error for NMF topic modeling across different pre-specified numbers of topics k, with $max\_df = 0.375$.



As shown in Figure 4.7, the reconstruction error was not informative for selecting k as no elbow is present in the trend, whereas topic coherence exhibited a peak at $k = 13$, leading the author to consider this value as a suitable choice. Although $k = 10$ or other

values of *max_df* yielded higher coherence scores, we ultimately selected the combination of $max\_df = 0.375$ and $k = 13$ because it led to more interpretable topics.

The assigned topics were validated on the 50-video sample by the human labelers, who attributed an integer score ranging from 0 to 4 to each video based on the assigned topics and their relative relevance. Across the entire sample, an average score of 3.32 was obtained, with 24 videos receiving a score of 4 and 20 videos receiving a score of 3. These results further support the effectiveness of the adopted topic modeling approach.

Table 4.9: Topics identified through NMF. For each topic, the table reports the top 10 stemmed words ranked by relevance, the label assigned by the author based on these keywords, and the number of videos in which the topic is the most relevant. The label *Metabolism* has been abbreviated to *Metab.* for display purposes.

Topic 0	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**
metabol	recip	drug	review	mounjaro	workout	surgeri
burn	add	medic	supplement	recip	leg	patient
sugar	cook	glp	offici	ingredi	knee	skin
reduc	cup	diabet	ingredi	tea	muscl	remov
drink	chicken	patient	websit	altern	walk	medic
blood	breakfast	obes	mitochondria	step	arm	lift
digest	salt	inject	metabol	drink	cardio	loos
boost	oil	studi	product	metabol	burn	danger
gut	onion	dose	guarante	method	minut	nan
insulin	prep	semaglutid	burn	sustain	core	doctor
Metab.	**Recipe**	**Medicine & Drugs**	**Supplement review**	**Mounjaro recipe**	**Workout**	**Surgery**
416	**384**	**407**	**201**	**70**	**167**	**161**

Topic 7	**Topic 8**	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**
hack	fast	yall	habit	fan	protein
trick	intermitt	walk	mindset	transform	carb
recip	hour	went	motiv	sister	gram
ice	window	she	tip	news	snack
salt	method	he	mind	star	yogurt
coffe	sugar	done	sleep	inspir	fiber
second	restrict	god	stress	she	egg
method	insulin	anyway	plan	reveal	sugar
simpl	salt	point	emot	fair	breakfast
burn	carb	man	step	appear	low
Trick	**Fasting**	**Personal storytelling**	**Mindset & Motivation**	**Transformation**	**Nutritional values**
69	**91**	**529**	**265**	**210**	**159**

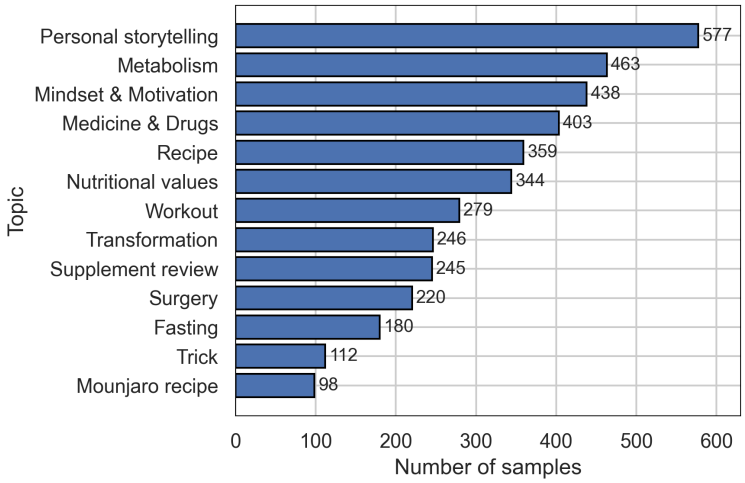
The topics selected are shown in Table 4.9. For each topic, the top 10 stemmed terms are reported together with the topic title assigned by the author. As can be observed, for

most topics the title coincides with the first listed term, since in these cases the first term effectively represents the overall semantic content of the remaining words, confirming its suitability as a label.

The most challenging topic to interpret was Topic 9, whose associated terms were rather general. In this case, we manually inspected randomly selected videos assigned to that topic, allowing common characteristics to emerge, namely the presence of individuals narrating their personal weight loss journeys and achievements. Based on this qualitative inspection, we labeled the topic *Personal storytelling*.

The *Personal storytelling* topic is also the most frequently assigned primary topic, as shown by the counts reported in the table. It is followed by *Metabolism* and *Medicine & Drugs*. The least frequent primary topics across all videos are *Trick* and *Mounjaro recipe*, likely due to their more specific focus. The *Trick* topic mainly characterizes videos in which specific weight-loss strategies or "tricks" are presented and explained.

Figure 4.8: Number of videos assigned to each topic based on the categorical feature representation.



Looking not only to the top topic for each video, but at the assigned topics based on threshold (see section 3.4), we obtain the counts shown in Figure 4.8. *Personal storytelling* confirms the first place for video count. *Mindset and Motivation* presents a count consistently higher then what observed earlier, indicating that this is often an important topic, even though not the primary one for most videos. The last positions in the video count rankings are occupied by about the same topics of before.

Overall, in addressing the first research question (Q1), we can conclude that the broad umbrella of diet and weight-loss content encompasses a wide variety of topics, ranging from physical exercise and recipes to medical information. A substantial proportion of the videos consist of first-person narratives, in which individuals share their weight-loss journeys, describing their challenges and achievements. Content related to the functioning of the body in relation to food and metabolism is also quite common. In addition,

some more specific topics clearly emerge, indicating the presence of well-defined content categories, such as *Trick*, *Mounjaro recipe*, and *Supplement review*.

The *Personal storytelling* topic showed the strongest correlations with several non-semantic variables. In particular, a Spearman correlation coefficient of $\rho_s = 0.58$ is observed with `duration_secs`, and $\rho_s = 0.30$ with `description_mentions`. Conversely, the correlation with `text_wordLen` is strongly negative ($\rho_s = -0.81$), suggesting that videos highly represented by the *Personal storytelling* topic are often characterized by shorter words.

The average word length, on the other hand, is positively correlated with the *Metabolism* topic ($\rho_s = 0.39$), while the *Supplement review* topic shows a positive correlation with `description_len` ($\rho_s = 0.35$).

To provide an additional sanity check for the topic modeling results, a correlation coefficient of $\rho_s = 0.47$ is observed between the *Recipe* topic and the *Food* channel category, confirming the semantic coherence of the extracted topics.

Finally, an interesting, albeit somewhat predictable, pattern emerges: all 125 videos associated with the keyword *mitolyn* during the exploratory phase of the study are assigned (at least) to the *Supplement review* topic, accounting for approximately half of the topic's total cardinality (245 videos). This pattern confirms the presence of coordinated or highly repetitive promotional content centered on this product, which significantly shapes the composition of the *Supplement review* topic.

4.3 Manual and LLM labeling

This section presents the outcomes of the video annotation process, mainly aimed at assessing information quality and the level of body-related content. Specifically, it details the establishment of a human-coded ground truth, the validation of the LLM's performance, and the subsequent application of the automated labeling to the entire dataset.

We first consider the annotations provided exclusively by the human labelers on the 50-video sample. The *agreement_scores* are used when available; otherwise, the *author_scores* are considered.

4.3.1 Human annotation

In Figure 4.9, the distribution of videos across the channel-characteristics variables is reported. Most channel owners are classified as *Individual*. In contrast, only a small number are categorized as *Commercial* or *Institution*. The relatively high frequency of the *Other* class suggests either that additional channel types could have been considered or that channel owners are too heterogeneous to be adequately captured by a small set of predefined categories.

A similarly high prevalence of the *Other* class is observed in the channel category statistics, leading to analogous considerations. Within this context, *Health* emerges as the most frequent category, followed by *Lifestyle*, which is consistent with the large number of videos assigned to the *Personal storytelling* topic by the NMF (4.2). The least represented

Figure 4.9: Number of sampled videos assigned to each channel owner type and channel category by the human labelers.

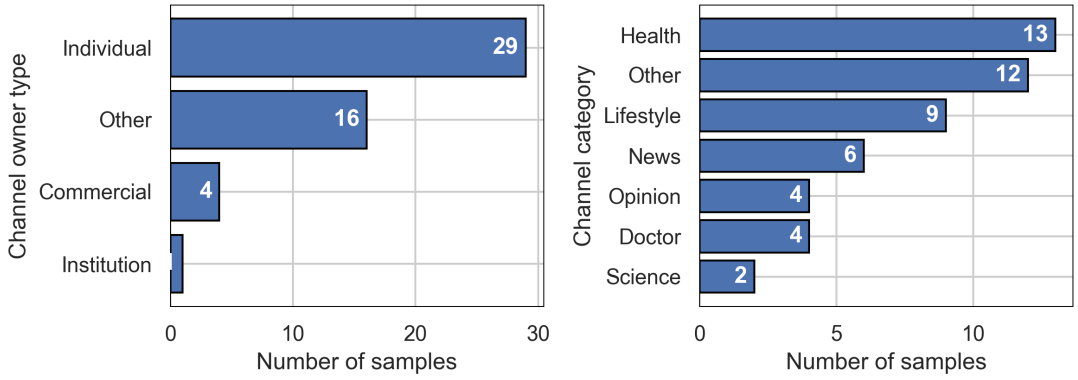
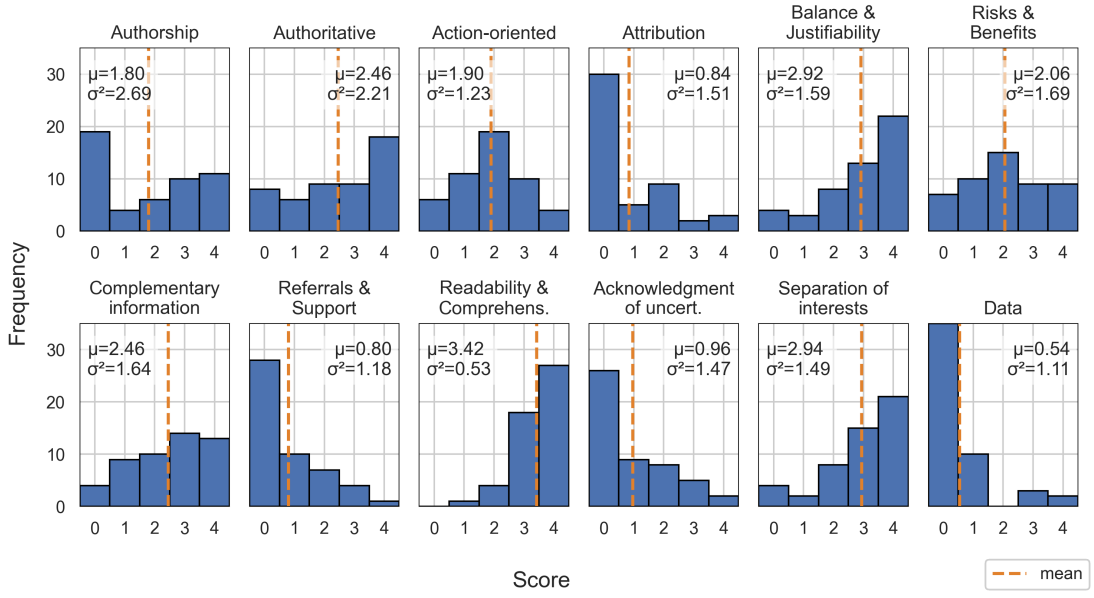


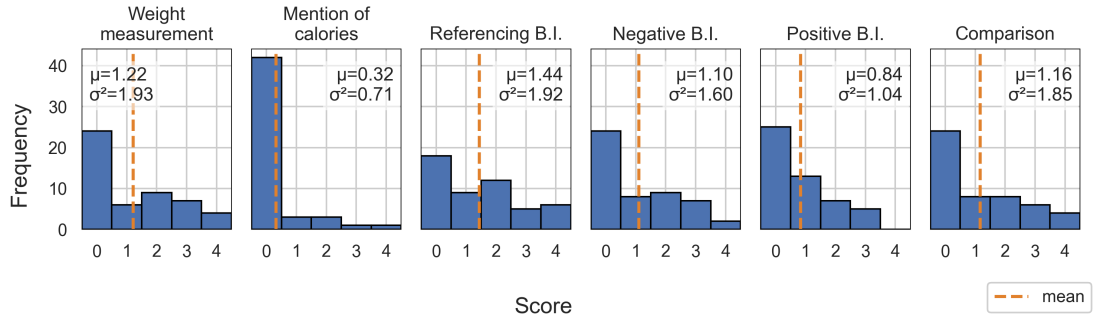
Figure 4.10: Distribution of the quality scores assigned by human labelers for each quality principle on the sampled videos, together with their mean (μ) and variance (σ^2).



categories are *Science* and *Doctor*, highlighting the relatively limited presence of scientific dissemination channels, even within diet- and weight-loss-related content.

Regarding the use of artificial intelligence in video production, 36 out of 50 videos were categorized by the labelers as *No AI tools were used in the video's production.* The second most frequent category, assigned to 8 videos, was *AI tools were used to modify or*

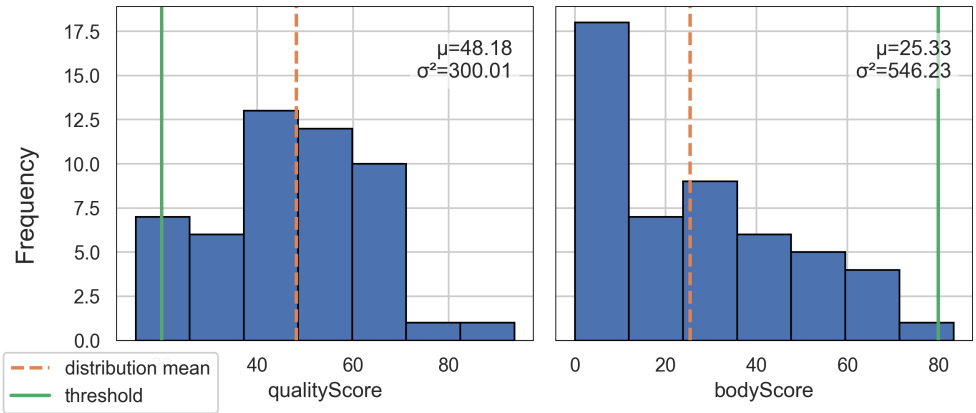
Figure 4.11: Distribution of the body scores assigned by human labelers for each body principle on the sampled videos, together with their mean (μ) and variance (σ^2).



generate a substantial portion of the video, significantly influencing its content, although the video was not entirely AI-generated. Based on what could be visually identified by human annotators, AI appeared to be mainly used for generating the background voice of the videos and for assembling images into collages or temporal sequences. However, the images themselves appeared to be real in almost all videos.

A more balanced distribution is observed with respect to the mention of brands or branded products: in 26 out of 50 videos, at least one such mention was identified by the annotators.

Figure 4.12: Distribution of the *quality score* and *body score*, expressed as percentages and computed from the human-assigned scores on the sampled videos. For each distribution, the mean (μ) and variance (σ^2) are reported. A green line indicates the risk threshold, set at a score of 20 for information quality and 80 for body-related content.



Figures 4.10 and 4.11 display the distributions of the scores assigned by the human labelers for each principle. As can be observed, only a few principles appear to follow an approximately normal distribution. Among the quality principles, *Attribution*, *Referrals*

& Support, *Acknowledgment of uncertainty*, and *Data* exhibit a pronounced right-skewed distribution. In contrast, *Balance & Justifiability*, *Readability & Comprehensibility*, and *Separation of interests* show a left-skewed pattern. The highest mean score is reported for *Readability & Comprehensibility*, whereas *Data* presents the lowest mean score, suggesting content highly understandable, but rarely supported by data.

Unlike the quality principles, all body-related principles display a right-skewed distribution, with 0 consistently representing the most frequent score. This suggests that most videos do not include verbal references to weight measurements, body image, or, in particular, calorie-related information. The same right-skewness is reflected in the overall *body score* distribution, as expected. By contrast, the *quality score* shows an approximately normal distribution, as illustrated in Figure 4.12.

Table 4.10: Agreement metrics between the author's annotations and those of the other human labelers on the whole double-annotated sample, reported for each principle of the two questionnaires, as well as for the *quality score* and *body score*. The latter are computed both as the average score rescaled to 100 (%) and as the average score rounded to the nearest integer (*round*). The reported metrics include the linear weighted Cohen's κ, the Brennan–Prediger κ, Spearman's ρ, and the corresponding p-values and Bonferroni adjusted p-values for Spearman's ρ. Confidence levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

	κ_w	κ_{BP}	MAE	ρ_s	p	p_{adj}
Authorship	0.295	8.500	1.400	0.398	0.082	
Authoritative	0.268	7.250	1.350	0.386	0.093	
Action-oriented	0.388	9.750	0.750	0.583	0.007	
Attribution	0.326	7.250	1.067	0.294	0.287	
Balance and justifiability	0.286	6.000	1.100	0.611	0.004	
Risks and benefits	0.394	6.000	0.842	0.683	0.001	*
Complementary information	0.364	11.000	1.100	0.474	0.035	
Referrals and support	-0.004	11.000	1.250	-0.118	0.622	
Readability and comprehens.	-0.092	12.250	0.650	-0.104	0.663	
Acknowledgment of uncert.	0.217	6.000	1.111	0.321	0.194	
Separation of interests	0.172	11.000	1.350	0.384	0.095	
Data	0.621	16.000	0.500	0.743	0.000	**
Weight measurement	0.469	12.250	0.900	0.798	0.000	***
Mention of calories	0.771	22.250	0.150	0.841	0.000	***
Referencing B.I.	0.396	8.500	1.105	0.553	0.014	
Negative B.I.	0.111	9.750	1.400	0.350	0.130	
Positive B.I.	0.333	11.000	0.750	0.404	0.077	
Comparison	0.338	9.750	1.050	0.645	0.002	*
quality score (%)	–	–	17.144	0.533	0.015	*
body score (%)	–	–	18.750	0.750	0.000	***
quality score (round)	0.455	12.250	0.600	0.596	0.006	*
body score (round)	0.301	8.500	0.800	0.625	0.003	**

These distributions should nevertheless be interpreted with caution, as they are based on a relatively small sample of the overall dataset.

In order to interpret the results more robustly, it was necessary to extend the labeling to the entire dataset using the LLM. This represents a sensitive step, as the LLM's ability to annotate the videos with a level of attention and interpretation comparable to that of a human annotator must first be assessed. To this end, we initially evaluated the agreement between different human labelers on the 20 double-labeled videos; the results are reported in Table 4.10.

As can be observed, only 5 out of the 18 principles show a statistically significant correlation between annotators, with *Weight measurement* and *Mention of calories* exhibiting the strongest agreement. This result is likely due to the subjective nature of most questions, whereas these two body-related principles, together with the quality principle *Data*, are comparatively more objective. Despite the limited significance at the individual principle level, the overall *quality score* and *body score* are significantly correlated across annotators. Spearman's correlation coefficient is 0.533 for the *quality score* and 0.750 for the *body score*, indicating a good level of association between human annotations.

When considering exact score agreement, Cohen's κ, computed on the rounded average scores, indicates moderate agreement for the *quality score* and fair agreement for the *body score*. The fact that Cohen's κ and Spearman's ρ provide partially divergent indications for quality and body-related content suggests that, although body scores assigned by one labeler are consistent with those of the other in terms of ranking, one annotator tends to assign systematically higher scores. This results in fewer exact score matches, a lower κ, and a higher error measure, as reflected by the *MAE* reported in the table.

Examining Cohen's κ at the level of individual principles, a substantial agreement is observed for *Mention of calories*. This is likely due to the high number of zero scores assigned to videos in which calories are not mentioned at all, thus reducing ambiguity and increasing consistency across labelers.

4.3.2 Model selection and Human-LLM agreement

Spearman's correlation coefficients and Cohen's κ values for each LLM configuration, with respect to human annotations on the validation set, are reported in Table 4.11. All combinations of temperature and prompting strategy yielded significant correlations between LLM and human annotations for both the overall *quality score* and *body score*. We obtained the highest Spearman's ρ for the *quality score* using temperature 0.1 with one-shot Chain-of-Thought (CoT) prompting, whereas for the *body score* we achieved the best performance with temperature 0 and zero-shot Chain-of-Thought prompting. However, when considering the time required to the LLM to annotate the 50 validation videos, CoT prompting proved substantially more demanding, requiring approximately four times the time needed for standard prompting. This makes its application to the full dataset considerably less practical.

A marked difference in input token usage across prompting strategies is also evident: zero-shot prompting requires approximately 400–500 input tokens per call, one-shot prompting around 1,400–1,800 tokens, and few-shot prompting approximately 2,600–3,000 tokens, excluding the tokens corresponding to the specific video details. Increasing the

Table 4.11: Agreement metrics (Spearman's ρ and linear weighted Cohen's κ) between the *agreement_scores* and the LLM annotations on the validation set, reported for each temperature value and prompting strategy. Spearman's ρ is computed on the average scores expressed as percentages, whereas Cohen's κ is computed on the rounded average scores. ZS, 1S, and FS denote zero-shot, one-shot, and few-shot prompting strategies, respectively. The Bonferroni adjusted p-values for Spearman's ρ are not reported, as all are below the 0.001 threshold. The last column reports the annotation time required by the LLM to process all 50 samples, in minutes.

	quality		**body**		**time**
	ρ_s	κ_w	ρ_s	κ_w	**(min)**
0, ZS	0.633	0.529	0.830	0.233	8:13
0, ZS CoT	0.629	0.446	0.871	0.304	24:19
0, FS	0.604	0.400	0.852	0.311	7:08
0, 1S CoT	0.602	0.326	0.848	0.360	29:31
0.1, ZS	0.644	0.507	0.834	0.277	6:55
0.1, ZS CoT	0.659	0.433	0.802	0.329	27:21
0.1, FS	0.644	0.418	0.829	0.329	6:47
0.1, 1S CoT	0.671	0.497	0.852	0.314	30:16

number of shots therefore results in a substantial rise in economic cost, which becomes even more pronounced in the case of CoT prompting, as output tokens must also be taken into account.

Given the higher time and economic costs, and considering that the correlation effect sizes are broadly comparable across configurations, we selected zero-shot standard prompting. Since temperature 0.1 yielded higher correlation coefficients than temperature 0 for both *quality score* and *body score*, we chose it as the final setting.

It is noteworthy that the observed correlation coefficients (approximately 0.6 for the *quality score* and 0.8 for the *body score*) are higher than those obtained when assessing agreement between human labelers. For a more detailed overview of agreement patterns, principle-level agreement metrics between LLM and human annotations are reported in Table 4.12.

Also in this case, although not all single-principle scores were significantly correlated, the overall *quality score* and *body score* showed significant and substantial correlation coefficients, in line with previous findings in the literature (Khalil et al. [2025]). Moreover, the agreement between human labelers and the LLM was overall higher than the agreement observed between different human labelers. This result further supports the effectiveness of the LLM in performing this type of annotation task and justifies its application to the full dataset.

Cohen's κ and Spearman's ρ again provide partially divergent indications for quality and body-related content, with an even larger discrepancy in Cohen's κ values computed on the rounded overall scores. This pattern suggests a systematic tendency of one party to assign higher scores on body-related principles. This interpretation is supported by Figure 4.13: in the scatter plot for the *body score*, the samples align closely with the

Table 4.12: Agreement metrics between the *agreement_scores* and the LLM annotations on the validation set, reported for each principle of the two questionnaires, as well as for the *quality score* and *body score*. The latter are computed both as the average score rescaled to 100 (%) and as the average score rounded to the nearest integer (*round*). The reported metrics include the linear weighted Cohen's κ, the Brennan–Prediger κ, Spearman's ρ, and the corresponding p-values and Bonferroni-adjusted p-values for Spearman's ρ. Significance levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

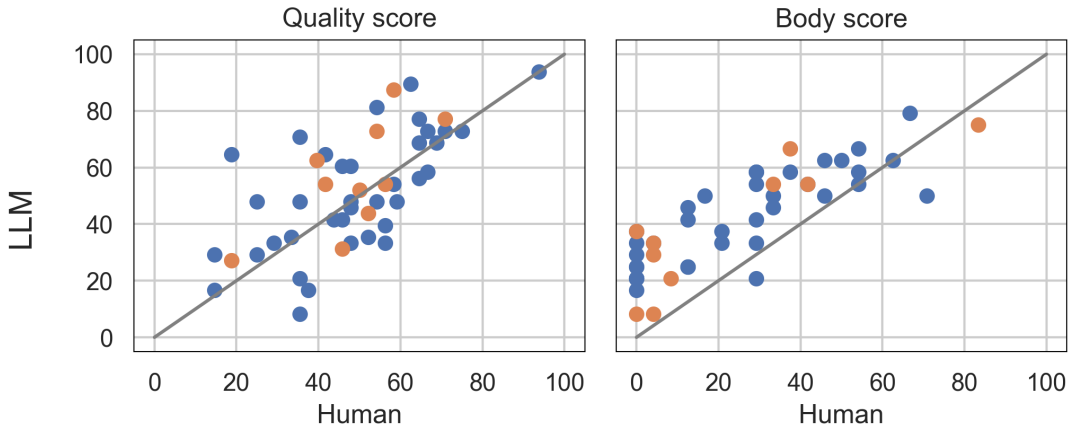
	κ_w	κ_{BP}	MAE	ρ_s	p	p_{adj}
Authorship	0.352	11.000	1.081	0.588	0.000	**
Authoritative	0.255	9.750	1.189	0.483	0.002	*
Action-oriented	0.298	11.000	0.865	0.682	0.000	***
Attribution	0.332	18.500	0.778	0.391	0.018	
Balance and justifiability	0.392	18.500	0.784	0.495	0.002	*
Risks and benefits	0.348	14.750	0.865	0.562	0.000	**
Complementary information	0.437	17.250	0.730	0.629	0.000	***
Referrals and support	0.322	21.000	0.784	0.441	0.006	
Readability and comprehens.	0.180	27.250	0.459	0.182	0.280	
Acknowledgment of uncert.	0.478	22.250	0.676	0.638	0.000	***
Separation of interests	0.100	17.250	1.027	0.085	0.617	
Data	0.272	17.250	0.784	0.190	0.260	
Weight measurement	0.182	7.250	1.622	0.827	0.000	***
Mention of calories	0.289	24.750	0.784	0.424	0.009	
Referencing B.I.	0.378	13.500	0.838	0.673	0.000	***
Negative B.I.	0.425	21.000	0.676	0.640	0.000	***
Positive B.I.	0.213	13.500	1.054	0.475	0.003	
Comparison	0.307	9.750	1.027	0.664	0.000	***
quality score (%)	-	-	12.295	0.644	0.000	***
body score (%)	-	-	18.468	0.834	0.000	***
quality score (round)	0.507	28.500	0.459	0.573	0.000	***
body score (round)	0.277	14.750	0.757	0.720	0.000	***

bisector trend, indicating strong rank correlation, yet they are predominantly located above the line, reflecting systematically higher scores from the LLM.

By contrast, in the scatter plot for the *quality score*, the points are more widely dispersed around the bisector, resulting in a lower correlation coefficient. However, they are more evenly distributed above and below the line, with a higher number of points lying directly on it. This leads to a lower MAE and a higher Cohen's κ, indicating better exact agreement despite weaker rank association.

Although the p-value is not a reliable indicator for the test set due to the small number of samples, the correlation effect sizes remain substantial, with $\rho_s = 0.669$ for the *quality score* and $\rho_s = 0.752$ for the *body score*. Test samples are shown in orange in Figure 4.13.

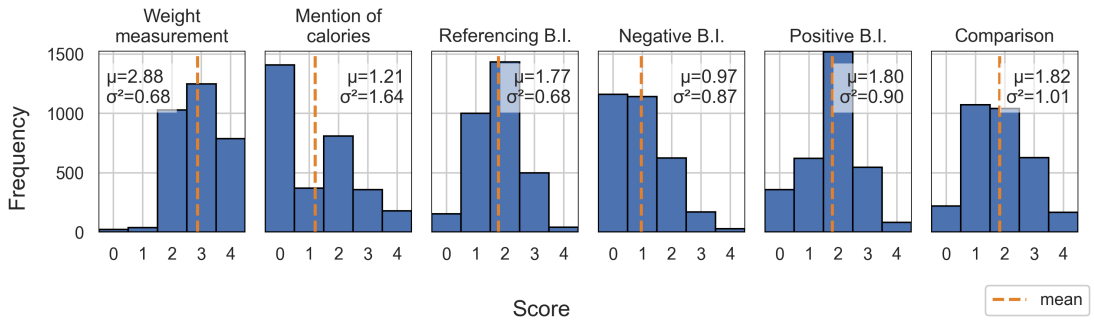
Figure 4.13: Comparison of the *quality score* and *body score* based on human annotations versus LLM annotations, for each sample in the validation set (blue) and test set (orange). The bisector is shown in gray in both scatter plots.



4.3.3 LLM-based annotation of the full dataset

We then applied the selected LLM model to annotate the entire dataset. The distributions of the scores obtained for each principle are shown in Figure 4.15 and Figure 4.14.

Figure 4.14: Distribution of the body scores assigned by the LLM for each body principle on the overall dataset, together with their mean (μ) and variance (σ^2).



Compared to the distributions derived from human labeling (Figure 4.10 and Figure 4.11), which were based on a smaller sample, the LLM-based distributions tend to appear more approximately normal. Nevertheless, several principles — including *Attribution*, *Readability & Comprehensibility*, *Separation of interests*, *Data*, and *Mention of calories* — remain skewed, and in the same direction as observed in the human-labeled sample. This confirms the high prevalence of easily comprehensible content that is often poorly supported by external references or data, as well as the relatively low presence of

calorie-related content or content characterized by hidden commercial interests.

Figure 4.15: Distribution of the quality scores assigned by the LLM for each quality principle on the overall dataset, together with their mean (μ) and variance (σ^2).

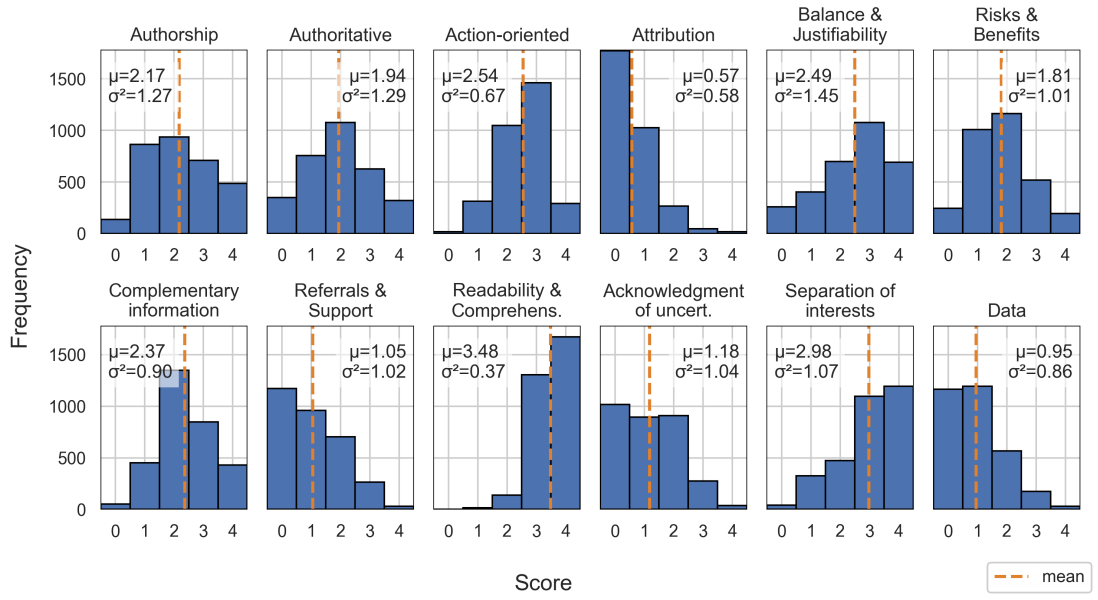
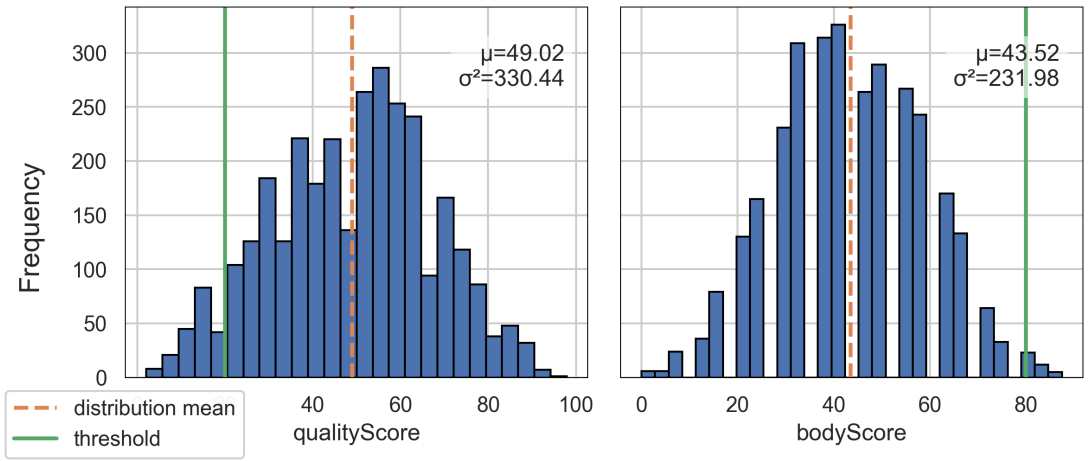


Figure 4.16: Distribution of the *quality score* and *body score*, expressed as percentages and computed from the LLM-assigned scores on the overall dataset. For each distribution, the mean (μ) and variance (σ^2) are reported. A green line indicates the risk threshold, set at a score of 20 for information quality and 80 for body-related content.



Finally, Figure 4.16 presents the distributions of the overall *quality score* and *body score*. In this case, the approximately normal shape of the distributions is more evident. The *body score* distribution is more concentrated toward lower values and only rarely exceeds the 80% threshold, probably conditioned by the absence of visual data.

4.4 Q2: Determinants of quality and body-related content

We now turn to the analyses conducted to address the second research question (Q2). We first examine the determinants of the *quality score* as the dependent variable, and subsequently replicate the analyses considering the *body score*.

4.4.1 Determinants of the *quality score*

This subsection investigates the relationship between the independent variables and the *quality score*. We begin by exploring univariate associations through correlation analyses and distributional comparisons, and then proceed to multivariate modeling using multiple linear regression to assess the joint effect of the predictors.

Univariate associations with numerical variables

Table 4.13 summarizes the Spearman's correlation coefficients obtained from the analysis of the relationships between numerical independent variables and the *quality score*. Topic variables are considered in their numerical representation. We log-transformed highly skewed variables prior to analysis.

Based on the Bonferroni-adjusted p-values, most non-semantic features show a statistically significant association with the outcome. The largest positive effect size is observed for `duration_secs (log)`, followed by the number of mentions in the description. Among the negatively correlated variables, `title_lenChar` and `text_wordLen` stand out, suggesting that longer titles and greater average word length are associated with lower-quality content. The negative association with average word length is consistent with the previously observed negative correlation between this variable and video duration.

By contrast, the number of hashtags, the ratio of exclamation marks, and description length do not exhibit a statistically significant association with the *quality score*. The relationships between the most strongly correlated non-semantic variables and the *quality score* are visually illustrated in Appendix B.

Turning to topic-related variables, several statistically significant associations are also observed. The strongest positive effect size is associated with *Medicine & Drugs*, followed by *Mindset & Motivation*. This result is consistent with expectations: the former topic is more closely related to medical content, which is more likely to be communicated by professionals or experts, while the latter typically involves motivational narratives rather than prescriptive claims or technical explanations, thereby reducing the likelihood of low quality.

Table 4.13: Spearman's correlation coefficient (ρ_s) between numerical independent variables (or their logarithmic transformations) and the *quality score*, together with the corresponding p-values (p) and Bonferroni-adjusted p-values (p_{adj}) intervals. Within each of the two sections of the table, variables are reported in descending order of the absolute values of the coefficients. Confid. levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

variable	ρ_s	p	p_{adj}
duration_secs (log)	0.413	3.034e-129	***
description_mentions (log)	0.248	3.381e-45	***
title_lenChar	-0.241	1.711e-42	***
text_wordLen	-0.234	2.633e-40	***
channelDescription_len	0.224	8.116e-37	***
description_links (log)	0.152	1.485e-17	***
channelVideoCount (log)	-0.149	6.433e-17	***
channelAge	0.143	9.163e-16	***
channelSubscriberCount (log)	-0.138	8.192e-15	***
title_desc_uppercaseRatio	-0.128	6.305e-13	***
title_desc_emojiRatio	-0.123	4.856e-12	***
description_hashtags (log)	-0.045	0.011	
title_desc_exlamRatio	-0.037	0.041	
description_len (log)	-0.015	0.414	
topic			
Medicine & Drugs	0.444	1.804e-151	***
Mindset & Motivation	0.265	1.357e-51	***
Personal storytelling	0.246	1.591e-44	***
Surgery	0.207	1.471e-31	***
Trick	-0.181	1.797e-24	***
Nutritional values	0.163	3.664e-20	***
Supplement review	-0.159	3.547e-19	***
Mounjaro recipe	-0.145	3.617e-16	***
Fasting	0.091	3.484e-07	***
Workout	0.087	1.031e-06	***
Recipe	-0.061	5.804e-04	*
Transformation	-0.049	0.007	
Metabolism	-0.006	0.720	

Conversely, strong negative correlations are observed for *Trick*, *Supplement review*, and *Mounjaro recipe*. These topics are more likely to involve non-expert creators presenting

scientific claims or health-related advice, increasing the risk of lower-quality information. The relationship between topic prevalence and *quality score* is visually represented in Figure B.4, where it is evident that greater association with *Medicine & Drugs* corresponds to higher quality scores, while the opposite trend is observed for negatively correlated topics.

Spearman's correlation coefficient is also statistically significant when examining the association between the level of AI use and the *quality score*, restricted to the 50-video subsample for which AI use was manually annotated. The estimated coefficient is $\rho_s = -0.466$, with a p-value of approximately 0.001, indicating that higher levels of AI-assisted content generation are associated with lower video quality.

Categorical variables: distributional comparisons

For categorical predictors, correlation coefficients cannot be computed directly. Therefore, we rely on Mann–Whitney U tests for binary comparisons and Kruskal–Wallis H tests

Table 4.14: Area under the ROC curve (AUC), p-value (p), Bonferroni-adjusted p-value intervals (p_{adj}), effect size r, and direction of the association for the Mann–Whitney U test conducted for each channel category with respect to the distribution of the *quality score*. A direction labeled as ↑ indicates that the distribution for the given category is significantly higher than that of all other videos, whereas ↓ indicates that it is significantly lower. Confidence levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

channel category	AUC	p	p_{adj}	r	direction
Knowledge	0.669	7.353e-26	***	0.188	↑
Politics	0.657	2.291e-09	***	0.107	↑
Society	0.653	3.803e-21	***	0.169	↑
Religion	0.619	0.001	*	0.059	↑
Hobby	0.588	0.021		0.041	
Health	0.581	1.824e-14	***	0.137	↑
Lifestyle (sociology)	0.569	1.398e-08	***	0.101	↑
Fashion	0.540	0.274		0.020	
Physical fitness	0.532	0.017		0.042	
Television program	0.521	0.319		0.018	
Food	0.497	0.845		-0.003	
Entertainment	0.374	1.666e-14	***	-0.137	↓
Film	0.266	3.013e-12	***	-0.125	↓
Music	0.194	7.964e-28	***	-0.195	↓
Pop music	0.176	9.947e-24	***	-0.179	↓
Independent music	0.098	3.870e-22	***	-0.173	↓
Music of Latin America	0.095	2.948e-23	***	-0.177	↓

when more than two groups are involved.

We begin with the channel category variable. The results of the Mann–Whitney tests are reported in Table 4.14. As expected, videos published by channels self-declared as *Knowledge* show a relatively high probability (AUC = 0.669) of having a higher *quality score* compared to videos from other channel categories. Significant differences in distributions are also observed for *Politics*, *Society*, and *Religion*, whose *quality score* distributions are higher than those of the complementary sets.

Conversely, negative effect sizes characterize music-related channel categories, further confirming the association between such categories and lower information quality. In particular, a video published by a channel categorized as *Music of Latin America* has only a 0.095 probability of exhibiting a higher *quality score* than a video from a channel belonging to a different category.

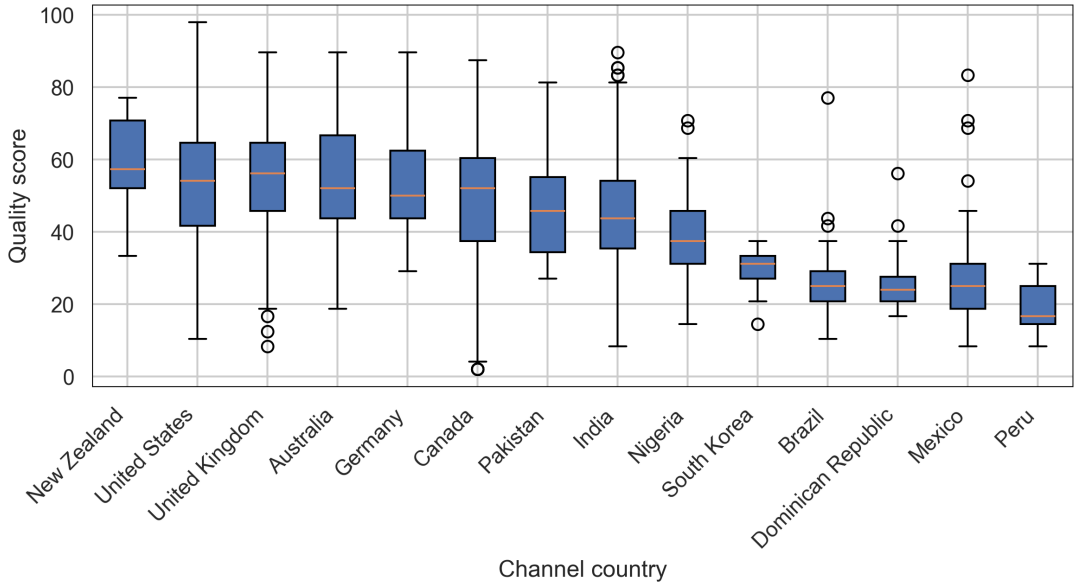
Although detailed results for video categories are not reported in tabular form, the Kruskal–Wallis test yields a statistic of $H = 283.49$ with a p-value below 0.001, confirming significant differences in *quality score* distributions across video categories. Subsequent Mann–Whitney tests indicate that the *Music* category is again associated with lower quality (AUC = 0.088), whereas *Science & Technology*, *News & Politics*, and *Education* are significantly associated with higher quality, as expected.

Table 4.15: Area under the ROC curve (AUC), p-value (p), Bonferroni-adjusted p-value intervals (p_{adj}), effect size r, and direction of the association for the Mann–Whitney U test conducted for each channel country with respect to the distribution of the *quality score*. A direction labeled as ↑ indicates that the distribution for the given country is significantly higher than that of all other videos, whereas ↓ indicates that it is significantly lower. Confidence levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

channel country	AUC	p	p_{adj}	r	direction
New Zealand	0.662	0.077		0.032	
United States	0.639	3.507e-41	***	0.240	↑
United Kingdom	0.605	8.098e-08	***	0.096	↑
Australia	0.575	0.081		0.031	
Germany	0.535	0.618		0.009	
Canada	0.489	0.640		-0.008	
Pakistan	0.462	0.533		-0.011	
India	0.435	0.009		-0.046	
Nigeria	0.348	0.022		-0.041	
South Korea	0.173	2.190e-07	***	-0.093	↓
Brazil	0.156	2.791e-09	***	-0.106	↓
Dominican Republic	0.141	2.828e-08	***	-0.099	↓
Mexico	0.138	3.653e-30	***	-0.204	↓
Peru	0.062	1.452e-08	***	-0.101	↓

To further examine these patterns, we consider the channel country variable. The Kruskal–Wallis test is again statistically significant. As shown in Table 4.15, several countries based in South or Central America are associated with lower *quality scores*. For instance, videos from Brazil have only a 0.156 probability of exceeding the *quality score* of videos from other countries, while for Peru this probability decreases to 0.062. These distributional differences are also clearly visible in Figure 4.17, where lower medians and narrower interquartile ranges can be observed for these countries.

Figure 4.17: Boxplots showing the distribution of the *quality score* across channel countries.



Interestingly, South Korea exhibits a similar pattern, with a significantly lower *quality score* distribution compared to the rest. By contrast, videos from channels based in the United States and the United Kingdom show probabilities slightly above 0.6 of having higher quality than the complementary sets. However, since these two countries account for a large share of the dataset, the comparison between each category and its complement may be partially driven by the lower quality associated with the remaining countries, including several South or Central American and less represented English-speaking countries.

Finally, we examine the categorical representation of topics, with results summarized in Table 4.16. The findings are consistent with previous analyses, confirming the positive association of *Medicine & Drugs* and *Mindset & Motivation* with higher *quality scores*. In contrast, *Trick*, *Mounjaro recipe*, and *Supplement review* are negatively associated with quality. For example, videos assigned to *Mounjaro recipe* have only a 0.144 probability of achieving a higher *quality score* than videos assigned to other topics.

Table 4.16: Area under the ROC curve (AUC), p-value (p), Bonferroni-adjusted p-value intervals (p_{adj}), effect size r, and direction of the association for the Mann–Whitney U test conducted for each topic with respect to the distribution of the *quality score*. A direction labeled as ↑ indicates that the distribution for the given topic is significantly higher than that of all other videos, whereas ↓ indicates that it is significantly lower. Confidence levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

topic	AUC	p	p_{adj}	r	direction
Medicine & Drugs	0.824	5.533e-98	***	0.375	↑
Mindset & Motivation	0.621	3.487e-16	***	0.146	↑
Personal storytelling	0.618	6.181e-19	***	0.159	↑
Surgery	0.612	2.799e-08	***	0.099	↑
Workout	0.556	0.002	*	0.056	↑
Nutritional values	0.544	0.008		0.047	
Fasting	0.516	0.474		0.013	
Metabolism	0.476	0.094		-0.030	
Recipe	0.459	0.011		-0.045	
Transformation	0.294	7.676e-27	***	-0.192	↓
Trick	0.227	7.246e-23	***	-0.176	↓
Mounjaro recipe	0.188	6.609e-26	***	-0.188	↓
Supplement review	0.144	6.127e-77	***	-0.332	↓

With regard to the categorical variables manually annotated on the 50-video subsample (namely channel owner type, channel category, and the presence of product or brand mentions), we did not detect any statistically significant difference in the Mann–Whitney U tests comparing the *quality score* distributions of each channel category with its complementary set, nor when comparing videos with and without product or brand mentions. Similarly, no significant differences emerged when contrasting the owner types *Institution* and *Commercial* with their respective complementary sets.

In contrast, the test assessing the *Individual* owner type yielded a statistically significant result, with a Bonferroni-adjusted p-value of 0.026, $AUC = 0.720$, and effect size $r = 0.372$. This indicates that a video posted by an *Individual* creator has a probability of 0.720 of achieving a higher *quality score* compared to videos posted by other owner types, including those classified under *Other*.

Although this finding may appear counterintuitive, it could be partially explained by the fact that creators producing lower-quality content are less easily identifiable, and may therefore be more likely to fall into the residual *Other* category.

Multiple linear regression models

Table 4.17 summarizes the regression coefficients for the models with *quality score* as the dependent variable, together with their corresponding significance levels.

Table 4.17: Regression coefficients and Bonferroni-adjusted p-value intervals for the multiple linear models with *quality score* as the dependent variable. Only variables with a statistically significant unadjusted p-value in at least one regression model were included. Separate models include numerical non-semantic variables (N), categorical non-semantic variables (C), topic variables (T), and their combinations. R^2_{adj} represents the adjusted coefficient of determination. Conf. levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

	N		**N+C**		**T**		**N+T**		**N+C+T**	
	$R^2_{adj} = 0.282$		$R^2_{adj} = 0.396$		$R^2_{adj} = 0.529$		$R^2_{adj} = 0.557$		$R^2_{adj} = 0.583$	
const	49.125	***	49.125	***	49.125	***	49.125	***	49.125	***
duration_secs	5.076	***	4.539	***			0.916		0.966	
channelSubscriberCount	-0.664		0.120				-0.100		0.184	
channelVideoCount	1.866	***	-0.338				0.043		-0.184	
channelAge	1.547	***	1.400	***			0.768	*	0.942	**
channelDescription_len	2.889	***	1.692	***			0.923	**	0.665	
description_len	-2.341	***	-0.725				1.560	***	1.494	***
title_lenChar	-3.213	***	-2.318	***			-0.856	*	-0.702	
text_wordLen	-0.089		-0.709				-0.889		-0.922	
title_desc_uppercaseRatio	-3.517	***	-3.002	***			-1.907	***	-1.757	***
title_desc_emojiRatio	-1.075	**	-0.688				-0.427		-0.407	
title_desc_exlamRatio	-0.713		-0.513				-0.348		-0.398	
description_links	2.186	***	1.425	***			0.299		0.224	
description_hashtags	-0.353		-0.161				-0.479		-0.276	
description_mentions	1.471	***	1.297	***			0.839	*	0.864	*
Entertainment			-0.659						-0.782	
Music			-0.877						-0.465	
News & Politics			1.482	*					0.331	
Pets & Animals			0.496						0.595	
Science & Technology			1.133	**					0.676	
(channel) Lifestyle (sociology)			-0.924						-0.532	
(channel) Health			1.824	***					1.184	*
(channel) Knowledge			2.616	***					1.396	***
(channel) Society			1.773	***					-0.366	
(channel) Entertainment			-2.045	***					-0.448	
(channel) Television program			1.235	*					0.763	
(channel) Film			-0.953						-0.909	*
(channel) Food			0.296						0.532	
(channel) Pop music			0.816						0.898	
(channel) Religion			-0.531						0.613	
United States			2.713	***					0.975	
United Kingdom			1.644	***					0.873	*
India			-0.189						-0.794	
Mexico			-1.190						-0.404	
Australia			0.779						0.325	
Brazil			-0.606						-0.075	
Pakistan			0.549						0.098	
Dominican Republic			-0.571						0.103	
Peru			-0.892						-0.440	
New Zealand			0.595						0.415	
Metabolism					1.727	***	1.373	***	1.219	**
Recipe					2.519	***	1.759	***	1.692	***
Medicine & Drugs					10.144	***	9.231	***	8.530	***
Supplement review					-2.299	***	-2.836	***	-2.565	***
Mounjaro recipe					-2.664	***	-2.639	***	-2.475	***
Workout					2.818	***	2.077	***	2.145	***
Surgery					3.215	***	2.698	***	2.319	***
Trick					-1.967	***	-2.090	***	-2.023	***
Fasting					1.268	***	1.059	***	0.984	**
Personal storytelling					4.532	***	2.930	***	2.855	***
Mindset & Motivation					5.135	***	4.169	***	4.059	***
Transformation					-1.213	***	-1.153	**	-0.783	
Nutritional values					2.079	***	1.602	***	1.396	***

We assessed multicollinearity through the Variance Inflation Factor (VIF), and we iteratively removed correlated features when necessary. When categorical variables were included, we excluded the video category *People & Blogs* due to severe multicollinearity. In the full model including all predictors, *People & Blogs* was removed with $VIF = 199.34$. After its exclusion, the highest remaining VIF was 4.497, associated with the channel category *Pop music*, indicating an acceptable level of multicollinearity.

From the table, it can be observed that adding additional predictors to the model reduces the statistical significance of several originally significant numerical non-semantic variables. Notably, topic variables alone explain a substantial portion of the variance in the outcome ($R^2_{adj} = 0.529$), exceeding the explanatory power of the combined non-semantic variables. This highlights the central role of topic-related features in predicting the *quality score*. The explanatory capacity of topics is further enhanced when they are combined with non-semantic predictors.

In the model including only numerical non-semantic variables, `duration_secs` exhibits the largest positive coefficient, indicating that longer videos are associated with higher quality. However, when topic variables are introduced, the magnitude of this coefficient decreases. This reduction can be attributed to the overlap between video duration and certain topics, particularly *Personal storytelling*, which are themselves correlated with `duration_secs`. Although we mitigated collinearity through VIF-based selection, predictors are not entirely independent.

In the final model including all retained variables, `channelAge`, `description_len`, and `title_desc_uppercaseRatio` remain statistically significant. The coefficients indicate that older channels are associated with higher-quality content and, interestingly, that a higher ratio of uppercase letters in titles and descriptions is negatively associated with quality. As expected, videos from channels categorized as *Knowledge* are associated with higher quality, whereas those from channels in the *Film* category are associated with lower quality. Channel categories such as *Society* and *Entertainment* lose significance in the full model, suggesting that their explanatory contribution is largely absorbed by topic variables.

Most topic coefficients remain highly significant in the final model. The largest positive coefficient is associated with *Medicine & Drugs*, consistent with previous analyses. Conversely, significantly negative coefficients are observed for the same topics previously identified as being linked to lower quality.

Overall, despite the inherent complexity of the task and the subjectivity typical of social science research, the final model captures a substantial proportion of the variance in the *quality score*.

4.4.2 Determinants of the *body score*

After examining the factors associated with the *quality score*, we now turn to the analysis of the determinants of the *body score*. As before, we begin with univariate analyses and then proceed to multivariate modeling.

Table 4.18: Spearman's correlation coefficient (ρ_s) between numerical independent variables (or their logarithmic transformations) and the *body score*, together with the corresponding p-values (p) and Bonferroni-adjusted p-values (p_{adj}) intervals. Within each of the two sections of the table, variables are reported in descending order of the absolute values of the coefficients. Conf. levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

variable	ρ_s	p	p_{adj}
text_wordLen	-0.316	2.210e-73	***
duration_secs (log)	0.311	2.707e-71	***
description_mentions (log)	0.143	9.724e-16	***
title_desc_uppercaseRatio	0.119	2.060e-11	***
title_desc_exlamRatio	0.074	3.292e-05	***
description_links (log)	0.059	9.274e-04	*
description_hashtags (log)	-0.026	0.142	
description_len (log)	0.018	0.312	
title_lenChar	0.018	0.313	
channelVideoCount (log)	-0.007	0.692	
channelAge	-0.006	0.737	
channelDescription_len	0.005	0.763	
channelSubscriberCount (log)	0.004	0.826	
title_desc_emojiRatio	-9.320e-04	0.958	
topic			
Personal storytelling	0.429	5.611e-140	***
Transformation	0.212	3.372e-33	***
Metabolism	-0.211	9.073e-33	***
Mindset & Motivation	0.200	1.485e-29	***
Workout	0.142	1.529e-15	***
Nutritional values	0.139	5.754e-15	***
Fasting	0.137	1.231e-14	***
Supplement review	-0.119	2.096e-11	***
Surgery	0.118	2.969e-11	***
Recipe	-0.085	1.791e-06	***
Medicine & Drugs	-0.046	0.010	
Trick	-0.035	0.049	
Mounjaro recipe	0.013	0.451	

Univariate associations with numerical variables

We first compute Spearman's correlation coefficients between numerical variables and the outcome. The results are reported in Table 4.18.

Compared to the *quality score*, a smaller number of variables, particularly non-semantic ones, show statistically significant associations with the *body score*. The strongest positive correlation is observed for the topic *Personal storytelling*. Consistently, `duration_secs (log)` is positively correlated with the outcome, while `text_wordLen` shows a negative association.

Textual emphasis indicators, such as the ratio of uppercase letters and the number of exclamation marks, are positively correlated with the level of body-related content. Among the other topic variables, the strongest positive associations with the *body score* are found for *Transformation*, *Workout*, and, somewhat unexpectedly, *Mindset & Motivation*.

Negative, although moderate, correlations are observed for topics such as *Metabolism*, *Recipe*, and *Supplement review*. This suggests that, although videos aligned with the *Supplement review* topic may exhibit lower informational quality in previous analyses, their primary focus is less explicitly centered on physical appearance and body-related content.

As in the previous analysis, the relationships between these numerical variables and the *body score* can be visually inspected through the scatter plots presented in Appendix B.

Finally, in contrast to the findings for the *quality score*, the level of AI use in video production, manually annotated for the 50-video subsample, does not exhibit a statistically significant correlation with the *body score*.

Categorical variables: distributional comparisons

Also when considering categorical variables, fewer significant relationships emerge compared to the analyses conducted for the *quality score*.

Videos belonging to the channel categories *Fashion* and *Entertainment* show probabilities of 0.670 and 0.654, respectively, of having a higher *body score* than videos in the complementary sets. This indicates a greater presence of body-related content within these categories. A distribution significantly shifted toward higher *body score* values is also observed for the channel category *Television program*.

Conversely, the categories *Society*, *Food*, and *Knowledge* are associated with lower levels of body-related content. In particular, a video from the *Knowledge* category has a probability of 0.362 of achieving a higher *body score* compared to videos from other categories. No statistically significant differences are found for the remaining channel categories.

A similar pattern emerges when video categories are examined. Only *Entertainment* and *People & Blogs* are significantly associated with higher *body scores*, whereas *How-to & Style*, *Education*, and *Science & Technology* are significantly associated with lower levels of body-related content. These findings are consistent with intuitive expectations.

Regarding channel countries, videos from the United States display a distribution of *body score* shifted toward higher values compared to the complementary set. However, this result may be influenced by the large size of this subgroup, as previously discussed. In contrast, significantly lower distributions are observed for India and South Korea. Notably, a video from South Korea has a probability of 0.210 of obtaining a higher *body score* relative to videos from other countries.

Table 4.19: Area under the ROC curve (AUC), p-value (p), Bonferroni-adjusted p-value intervals (p_{adj}), effect size r, and direction of the association for the Mann–Whitney U test conducted for each topic with respect to the distribution of the *body score*. A direction labeled as ↑ indicates that the distribution for the given topic is significantly higher than that of all other videos, whereas ↓ indicates that it is significantly lower. Confidence levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

topic	AUC	p	p_{adj}	r	direction
Transformation	0.744	3.100e-37	***	0.227	↑
Personal storytelling	0.726	3.092e-65	***	0.304	↑
Surgery	0.608	9.184e-08	***	0.095	↑
Fasting	0.602	3.519e-06	***	0.083	↑
Mindset & Motivation	0.594	2.309e-10	***	0.113	↑
Nutritional values	0.560	2.925e-04	**	0.065	↑
Mounjaro recipe	0.550	0.091		0.030	
Trick	0.514	0.612		0.009	
Workout	0.489	0.532		-0.011	
Medicine & Drugs	0.402	2.027e-10	***	-0.113	↓
Recipe	0.369	3.806e-16	***	-0.145	↓
Supplement review	0.354	2.066e-14	***	-0.136	↓
Metabolism	0.347	5.444e-26	***	-0.188	↓

Topic-based comparisons reveal more pronounced differences in distributions, as shown in Table 4.19. The topics *Transformation* and *Personal storytelling* confirm their strong positive association with higher *body scores*, followed by *Surgery*, *Fasting*, and *Mindset & Motivation*. Conversely, lower levels of body-related content are associated with the topics mentioned in the previous section. Interestingly, *Medicine & Drugs* exhibits a significant difference in distribution despite not showing a significant correlation with the outcome in the numerical representation.

Finally, we found no statistically significant differences in the *body score* distributions when comparing the groups defined by the human-annotated categorical variables within the 50-video subsample, neither the channel owner type, the channel category or the mention of brands or branded products.

Multiple linear regression models

We now examine the results of the multiple linear regression models with *body score* as the dependent variable, summarized in Table 4.20. We assessed multicollinearity through the Variance Inflation Factor (VIF), and the only variable that we removed was the video category *People & Blogs* when categorical predictors were included.

Table 4.20: Regression coefficients and Bonferroni-adjusted p-value intervals for the multiple linear models with *body score* as the dependent variable. Only variables with a statistically significant unadjusted p-value in at least one regression model were included. Separate models include numerical non-semantic variables (N), categorical non-semantic variables (C), topic variables (T), and their combinations. R^2_{adj} represents the adjusted coefficient of determination. Conf. levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

	N		**N+C**		**T**		**N+T**		**N+C+T**	
	$R^2_{adj} = 0.122$		$R^2_{adj} = 0.206$		$R^2_{adj} = 0.346$		$R^2_{adj} = 0.379$		$R^2_{adj} = 0.394$	
const	43.492	***	43.492	***	43.492	***	43.492	***	43.492	***
duration_secs	1.879	***	1.809	***			-2.565	***	-2.435	***
channelSubscriberCount	0.498		0.325				0.604		0.482	
channelVideoCount	-1.113	**	-1.213	*			-1.029	**	-0.557	
channelAge	-1.202	***	-0.821				-0.947	**	-0.711	
channelDescription_len	-0.607		-0.365				-0.704		-0.742	
description_len	1.048	*	1.356	**			1.789	***	1.752	***
title_lenChar	0.905	*	0.538				0.317		0.320	
text_wordLen	-4.314	***	-4.481	***			-3.351	***	-3.265	***
title_desc_uppercaseRatio	0.336		0.098				0.480		0.347	
description_hashtags	-0.734		-0.123				-0.359		-0.142	
Autos & Vehicles			-0.548						-0.432	
Education			-1.045	**					-0.589	
Entertainment			-0.215						-0.591	
How-to & Style			-1.057	**					-0.611	
News & Politics			-0.253						-1.023	
Sports			-1.376	***					-0.732	
(channel) Lifestyle (sociology)			-0.985						-0.422	
(channel) Health			1.456	*					1.320	**
(channel) Food			-1.575	***					-0.417	
(channel) Knowledge			-1.077	**					-0.777	
(channel) Society			-0.772						-0.412	
(channel) Entertainment			3.229	***					0.790	
(channel) Television program			-0.276						-0.681	
(channel) Film			-0.670						0.130	
(channel) Religion			-0.745						-0.944	*
(channel) Fashion			0.856	*					0.195	
United States			0.840						0.100	
Canada			0.531						0.517	
Australia			0.635						0.273	
South Korea			-1.077	**					-0.491	
Metabolism					0.104		1.205	***	1.031	*
Medicine & Drugs					0.657		2.052	***	2.194	***
Supplement review					1.494	***	1.388	**	0.797	
Mounjaro recipe					2.400	***	2.414	***	2.082	***
Workout					1.270	***	1.661	***	1.592	***
Surgery					2.089	***	2.415	***	2.380	***
Trick					2.035	***	1.899	***	1.667	***
Fasting					2.005	***	2.269	***	2.199	***
Personal storytelling					7.591	***	7.805	***	7.812	***
Mindset & Motivation					3.361	***	4.145	***	4.130	***
Transformation					5.407	***	6.220	***	5.901	***
Nutritional values					2.849	***	2.949	***	2.822	***

As in the case of the *quality score*, the inclusion of topic variables substantially increases the proportion of explained variance. Nevertheless, the maximum R^2_{adj} achieved in this setting remains lower than that obtained for the *quality score*. Notably, all coefficients associated with topic variables are positive. This indicates that videos more strongly associated with one or more identified topics tend to exhibit higher levels of body-related content, whereas videos that are weakly characterized by these topics tend to have lower *body scores*. The topic *Recipe* is the only one that never reaches statistical significance across the different model specifications.

The largest topic coefficient is associated with *Personal storytelling*, once again confirming its strong positive relationship with body-related content. Consistently, associations with `duration_secs` and `text_wordLen` remain observable, reflecting the overlap between narrative-style content and video length or linguistic structure. A positive association is also found for `channelDescription_len`, which becomes even more statistically significant in the full model.

Among channel categories, *Health* shows the strongest positive association with the *body score*, a result that may raise questions regarding the type of health-related content being produced. Conversely, as expected, videos from channels categorized under *Religion* are associated with lower levels of body-related content.

Although the overall explained variance is lower than in the previous analysis, it remains sufficiently high to support predictive modeling. Moreover, the fact that the observed relationships are largely consistent with intuitive expectations provides additional support for the validity of the topic modeling procedure and the subsequent labeling process.

However, since we only considered textual data, body references expressed exclusively through visual content are not captured. As a result, the body-related scores likely represent an underestimate of the actual presence of body-related references in the videos.

The strong correlation between the *body score* and the *Personal storytelling* topic may also be biased by the fact that, in videos aligned with this topic, body-related content is more often explicitly stated in the narration rather than conveyed visually.

4.5 Q3: Quality, body-related content, and user engagement

After addressing the second research question, we now turn to the third (Q3), which concerns the relationships between user engagement and the other variables considered in the study.

To compute the engagement rate, we restricted the analysis to videos with non-null values for view count, like count, and comment count measured 70 days after upload, and with a non-zero view count. This filtering step reduced the sample size to 2,614 videos.

4.5.1 Univariate analyses

As in the previous research question, we begin with univariate analyses. Spearman's correlation coefficients were computed between the two independent variables, *quality*

score and *body score*, and the two dependent variables, view count and engagement rate.

After Bonferroni adjustment, all four correlations are statistically significant, with coefficients summarized in Table 4.21. Although the magnitudes of the coefficients are modest, several patterns emerge. Higher levels of body-related content are associated with both higher view counts and higher engagement rates. In contrast, the *quality score* is negatively correlated with view count but positively correlated with engagement rate. This suggests that lower-quality videos tend to receive more views overall, whereas higher-quality videos generate relatively more interaction (likes and comments) among viewers.

Table 4.21: Spearman's correlation coefficients per combination of variables.

quality score view count	**body score view count**	**quality score engagement rate**	**body score engagement rate**
-0.053	0.157	0.059	0.112

Despite their statistical significance, these correlations are small in magnitude and should therefore be interpreted with caution.

Furthermore, the observed correlations could be mediated by video topics or metadata. In fact, the *Personal storytelling* topic is significantly positively correlated with both the *body score* and engagement metrics, showing a correlation coefficient of $\rho_s = 0.208$ with view count and $\rho_s = 0.319$ with engagement rate. Additionally, several non-semantic variables previously identified as strongly correlated with *Personal storytelling* now exhibit analogous correlation significance and direction with engagement metrics. This suggests that higher view counts and engagement rates may be more strongly associated with the *Personal storytelling* topic than with the level of body-related content itself.

Similar patterns with the *quality score* are less evident. However, some notable significant correlations emerge between view count and the ratio of uppercase letters in titles and descriptions ($\rho_s = 0.279$), as well as with the ratio of emojis ($\rho_s = 0.139$). This suggests that videos using these stylistic elements to increase visibility may indeed achieve higher exposure. At the same time, no significant correlation is observed between view count and the use of exclamation marks. An alternative interpretation of these correlations is that repurposed music channels, which already had access to a wide audience, were later populated with videos featuring emojis and uppercase titles and descriptions, while continuing to benefit from their pre-existing visibility.

After *Personal storytelling*, the topics with the strongest absolute correlations with view count are *Mindset & Motivation* and *Metabolism*, with negative coefficients of $\rho_s = -0.207$ and $\rho_s = -0.141$, respectively. This suggests that videos more strongly associated with these topics tend to receive fewer views.

Metabolism is also the second most significantly correlated topic (after *Personal storytelling*) with engagement rate, again showing a negative coefficient ($\rho_s = -0.126$). It is followed by *Medicine & Drugs*, with $\rho_s = -0.118$, suggesting that audiences engage less with content that is more closely related to scientific or medical discourse.

Finally, the ratio of uppercase letters and emojis in titles and descriptions, and, in this case, also the ratio of exclamation marks, shows a significant positive correlation with engagement rate. This indicates that such stylistic features are not only associated with

larger audiences but also with higher levels of viewer interaction.

To further explore the relationships between engagement metrics and quality or body-related content levels, Table 4.22 and Table 4.23 report the correlation coefficients for each individual quality and body principle, together with their corresponding p-values. However, since not all the scores assigned by the LLM to individual principles in the 50-video sample were significantly correlated with those assigned by human annotators, it is difficult to fully assess the model's ability to provide reliable scores at the single-principle level. Therefore, the following results should be interpreted with caution.

Table 4.22: Spearman's correlation coefficient (ρ_s) between single-principle scores and the view count, together with the corresponding p-values (p) and Bonferroni-adjusted p-values (p_{adj}) intervals. Within each of the two sections of the table, principles are reported in descending order of the absolute values of the coefficients. Confidence levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

principle	ρ_s	p	p_{adj}
Separation of interests	-0.202	5.618e-26	***
Balance and justifiability	-0.115	3.033e-09	***
Risks and benefits	-0.100	2.154e-07	***
Authorship	0.068	4.765e-04	**
Authoritative	-0.056	0.004	
Attribution	0.051	0.008	
Action-oriented	-0.048	0.014	
Complementary information	-0.034	0.078	
Readability and comprehens.	-0.031	0.114	
Data	-0.009	0.639	
Acknowledgment of uncert.	-0.008	0.684	
Referrals and support	-0.003	0.861	
Negative B.I.	0.151	4.670e-15	***
Weight measurement	0.139	7.242e-13	***
Referencing B.I.	0.137	1.535e-12	***
Comparison	0.128	3.606e-11	***
Mention of calories	0.065	8.748e-04	*
Positive B.I.	-0.010	0.599	

Focusing first on the body-related portion of the questionnaire, nearly all principles show a positive correlation with view count, indicating that references to the body, particularly negative or critical ones, are associated with higher numbers of views. A similar, though less widespread, pattern is observed for engagement rate, where fewer principles

exhibit statistically significant correlations.

Table 4.23: Spearman's correlation coefficient (ρ_s) between single-principle scores and the engagement rate, together with the corresponding p-values (p) and Bonferroni-adjusted p-values (p_{adj}) intervals. Within each of the two sections of the table, principles are reported in descending order of the absolute values of the coefficients. Confidence levels: $p_{adj} < 0.001$ ***, $p_{adj} < 0.01$ **, $p_{adj} < 0.05$ *.

principle	ρ_s	p	p_{adj}
Complementary information	0.133	7.513e-12	***
Authorship	0.124	2.252e-10	***
Readability and comprehens.	0.117	2.217e-09	***
Referrals and support	0.087	9.268e-06	***
Action-oriented	0.082	2.813e-05	***
Data	-0.079	5.390e-05	***
Acknowledgment of uncert.	0.053	0.006	
Authoritative	0.041	0.038	
Attribution	-0.032	0.103	
Balance and justifiability	0.029	0.138	
Separation of interests	-0.019	0.338	
Risks and benefits	-0.017	0.392	
Weight measurement	0.149	1.915e-14	***
Positive B.I.	0.114	4.455e-09	***
Comparison	0.066	6.916e-04	*
Mention of calories	0.060	0.002	*
Referencing B.I.	0.028	0.159	
Negative B.I.	0.020	0.318	

Consistently with the results obtained for the overall *quality score*, opposite patterns emerge when comparing view count and engagement rate at the level of individual quality principles. For view count, the strongest significant correlations are negative and involve *Separation of interests*, *Balance and justifiability*, and *Risks and benefits*. This indicates that videos that are more balanced, clearly explain risks and benefits, and maintain independence from potential external interests tend to receive fewer views.

Conversely, when engagement rate is considered, the most significant coefficients are positive and are associated with *Complementary information*, *Authorship*, *Readability and comprehensibility*, and *Referrals and support*. Notably, these principles differ from those associated with view count, suggesting that view count and engagement rate capture distinct dimensions of content quality. In this context, engagement appears to be more strongly related to clarity, transparency, and the provision of supportive or complementary

information. Greater transparency regarding authorship, clearer presentation of information, and the inclusion of disclaimers or references to professional support are associated with higher engagement rates.

One possible interpretation is that viewers are more inclined to interact with content that they perceive as clear, transparent, and professionally grounded. Alternatively, such videos may stimulate more discussion, thereby increasing the number of comments.

Finally, it is important to emphasize that these findings are purely correlational. Given the presence of numerous potential confounding factors, no causal conclusions can be drawn from these analyses.

4.5.2 Multiple linear regression

We estimated two multiple linear regression models, using the individual principle scores as independent variables. In the first model, we considered view count as the dependent variable, while in the second model the dependent variable was engagement rate.

The model with view count as the outcome did not yield any statistically significant coefficients, preventing any conclusions regarding the independent contribution of individual quality principles to the number of views within this multivariate framework.

In contrast, the model with engagement rate as the dependent variable produced two statistically significant coefficients, equal to 0.0137 and 0.0116, corresponding to *Complementary information* and *Positive portrayal of body image*, respectively. However, given the very small magnitude of these coefficients, their practical relevance appears limited. Therefore, drawing substantive conclusions based on these effect sizes would not be appropriate.

4.6 Q4: Predicting risk-related scores

To address our final research question (Q4), we employed standard neural networks, specifically multi-layer perceptrons, to assess their ability to detect low-quality or highly body-centered videos based on video topics and non-semantic metadata. The performances achieved by the selected models (through cross-validation) to predict the *quality score* for each model configuration are summarized in Table 4.24, for both the training and test sets, while their selected hyperparameters are reported in Table 4.25.

Overall, the results in the prediction of both *quality score* and *body score* are rather disappointing, likely due to the strong imbalance in the distribution of the output classes. In particular, across the entire dataset, only 199 out of 3129 samples present an LLM-assigned *quality score* below 20 and only 17 samples have an LLM-assigned *body score* above 80. In such conditions, models that always predict the negative class (high quality or low body-centered) can achieve very high accuracy, as can indeed be observed in our results.

When focusing on the prediction of quality, precision is generally high for similar reasons: when only a very small number of videos are predicted as low-quality, it becomes more likely that those videos are indeed low quality. The metric of primary interest in this analysis is recall, which measures the proportion of actual positive samples that are

Table 4.24: Performance metrics obtained by the re-trained CV selected models on the training and test sets for the prediction of the *quality score*, in their original form (Regr.) and binarized form (Classif.). Regression metrics include mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R^2. Classification metrics instead include accuracy, precision, recall, and F1-score; in the regression setting, these are computed after binarizing the predicted scores.

		MSE	**RMSE**	**MAE**	R^2	**Acc.**	**Prec.**	**Recall**	**F1**
Classif.	train	-	-	-	-	0.991	0.929	0.923	0.926
	test	-	-	-	-	0.960	0.788	0.591	0.675
Regr.	train	57.732	7.598	5.783	0.823	0.966	0.829	0.561	0.669
	test	128.659	11.343	8.603	0.626	0.955	0.786	0.500	0.611

Table 4.25: Hyperparameters of the best-performing model selected through cross-validation for predicting the *quality score*, in their original form (Regr.) and binarized form (Classif.).

	N. layers	**Layer size**	`activation`	`alpha`	`batch_size`
Classif.	10	50	relu	0.05	10
Regr.	1	200	relu	0.05	auto

correctly identified as positive. Considering this metric, the obtained results are mediocre. In the classification task, recall decreases substantially between the training and test sets, suggesting potential overfitting, and it shows even lower values when computed on the binarized predictions of the regression models.

By contrast, the regression model without binarization exhibits considerably better results. In particular, the R^2 obtained on the test set is higher than that achieved by the multiple linear regression model (see section 4.4.1), indicating a higher explained variance and better predictive performance when the absolute *quality score* is considered. On the test set, the model produced an average absolute error of 8.603 on a scale ranging from 0 to 100, indicating reasonably acceptable predictive accuracy.

Conversely, when the prediction concerns the level of body-related content, the *MAE* is considerably higher and the R^2 is negative, indicating very poor predictions, even worse than those obtained by a model that always predicts the mean value of the dependent variable. The multi-layer perceptrons adopted therefore do not appear to be suitable for predicting the *body score*, suggesting that multiple linear regression provides higher predictive power in this context. Predictive performance deteriorates further when binary outputs are considered, with the classification model likely overfitting and the regression model with binarization achieving very low performance metrics on both the training and test sets.

Chapter 5

Conclusion

This thesis investigated the characteristics of diet and weight-loss content on YouTube, with particular attention to the quality in the information spread, the presence of body-related discourse, and their relationship with user engagement. By combining data collection from the YouTube platform, topic modeling techniques, human annotation, and large language model (LLM)–assisted labeling, we constructed a large-scale dataset that enabled a systematic analysis of these phenomena.

The study addressed four main research questions. First, we explored the main thematic structures of dieting and weight-loss videos through topic modeling. Second, we analyzed the variables associated with information quality and body-related content, considering both semantic variables derived from topics and non-semantic metadata. Third, we examined the relationship between information quality, body-related content, metadata, and user engagement metrics. Finally, we developed standard predictive models to facilitate the identification of low-quality or strongly body-centered content, which may pose higher risks to viewers in terms of body image disturbance or the potential development of disordered eating behaviors and eating disorders.

The preliminary exploration of the dataset revealed several notable patterns regarding content distribution and user engagement. First, both video duration and user interaction metrics (views, likes, and comments) exhibited highly right-skewed distributions. This indicates that while the vast majority of videos are relatively short and receive limited interaction, a small subset of videos captures a disproportionately large share of audience attention.
Regarding channel characteristics, content production in the dataset is predominantly localized in the United States according to the self-declared channel location, with *Health*, *Lifestyle*, and *Physical fitness* emerging as the most frequent self-declared channel categories overall.
A keyword-based investigation revealed further anomalous and potentially manipulative content patterns. Specifically, videos promoting the dietary supplement *Mitolyn* and a recipe for a natural version of the diabetes medication *Mounjaro* were surprisingly concentrated within the *Music* category and were predominantly uploaded by channels based

in South or Central America, particularly in Mexico and Brazil. Manual inspection suggested that these channels were often repurposed music channels operated by a coordinated group of individuals, sharing the same actors across different accounts to promote pharmaceutical products or suggest alternative recipes.

With respect to the first research question (Q1), the application of topic modeling revealed a heterogeneous thematic landscape in dieting and weight-loss videos. The identified topics cover both informational and experiential dimensions of the discourse, ranging from medically oriented content to personal narratives and practical advice. In particular, topics such as *Medicine & Drugs* and *Metabolism* reflect videos that focus on medical explanations or pharmacological aspects of weight loss, while topics such as *Personal storytelling*, *Transformation*, and *Mindset & Motivation* highlight the narrative and motivational dimension of the platform. Other topics, including *Supplement review*, *Trick*, and *Mounjaro recipe*, capture more specific forms of content often centered on practical suggestions or individual strategies for weight loss.

Preliminary to the investigation of the second research question, in order to scale the evaluation of information quality and body-related content to the full dataset, we tested a GPT model under several prompting strategies. We ultimately selected zero-shot standard prompting with a temperature of 0.1, as it achieved a strong rank correlation with human annotations, exceeding even the inter-rater agreement among the human labelers, while remaining computationally and economically efficient. Applying this configuration to the entire dataset produced approximately normal distributions for both the overall *quality score* and the *body score*. The automated labeling confirmed the widespread presence of highly comprehensible content that is often poorly supported by data or external references, whereas strongly explicitly body-focused material appeared to be relatively rare.

To address the second research question (Q2), we then investigated the determinants of information quality and body-related content. The analyses showed that both semantic and non-semantic variables contribute to explaining variations in these outcomes, although their roles differ across the two dimensions.

With regard to information quality, topic-related variables emerged as particularly relevant predictors. Videos associated with medically oriented or informational topics, such as *Medicine & Drugs*, tend to achieve higher quality scores, likely reflecting the presence of more structured explanations and references to medical knowledge. Conversely, topics such as *Trick*, *Supplement review*, and *Mounjaro recipe* are associated with lower quality scores, suggesting that these types of content may more frequently include simplified, anecdotal, or potentially misleading information. Among non-semantic variables, several metadata characteristics also showed significant associations with quality. In particular, longer video duration and older channel age were positively related to quality, while stylistic features such as higher proportions of uppercase letters in titles or descriptions were negatively associated with quality. Additionally, channel categories related to knowledge-oriented content were linked to higher quality, whereas entertainment or music-related categories tended to correspond to lower quality levels, probably due to the aforementioned suspicious patterns.

When considering body-related discourse, the determinants showed a partially different structure. Topics emphasizing personal experience or visible changes in physical

appearance, such as *Personal storytelling*, *Transformation*, and *Workout*, were positively associated with higher levels of body-related content. These topics often involve narratives centered on individual journeys, physical changes, or motivational framing, which naturally verbally emphasize the body as a central element of the discourse. In contrast, topics such as *Metabolism*, *Recipe*, or *Supplement review* showed weaker or negative associations with body-related content, suggesting a more informational or practical orientation rather than a focus on explicit physical appearance. The regression analyses confirmed the importance of topic variables also in this case, although the overall proportion of explained variance was lower than for the quality score, indicating that body-related discourse may depend on a broader set of contextual factors or was difficult to capture through the LLM-assisted labeling, without the support of video images.

The third research question (Q3) explored the relationship between the risk-related scores and user engagement, as well as the relationship between engagement and numerical metadata variables. The analyses revealed that body-related discourse tends to be positively associated with both view counts and engagement rates, indicating that videos emphasizing body-related themes generally have higher levels of audience attention and interaction. However, this relationship may be mediated by the *Personal storytelling* topic, which shows strong correlations with both the *body score* and engagement metrics. In contrast, information quality showed a more nuanced relationship with engagement metrics. While higher quality was associated with slightly higher engagement rates, it was negatively correlated with view counts, suggesting that videos providing more balanced, evidence-based information do not generally reach the largest audiences. These findings highlight a potential tension between informational quality and visibility on the platform, where content that is more sensational, personal, or visually oriented may achieve greater reach. Engagement patterns appear to be more strongly associated with stylistic metadata and certain video topics than with the risk-related scores themselves. In particular, stylistic elements such as uppercase letters and emojis in titles and descriptions are positively correlated with both view count and engagement rate, while topics related to scientific or medical discourse (e.g., *Metabolism* and *Medicine & Drugs*) show negative correlations with user engagement.

To address the fourth research question (Q4), we employed multi-layer perceptrons to evaluate whether video topics and non-semantic metadata could be easily and cheaply used to predict low-quality or highly body-centered videos, both in a regression and classification settings. Overall, the obtained results were mixed and generally limited. While the NN regression model achieved acceptable performance in predicting the continuous *quality score*, with a reasonably low *MAE* and higher R^2 than multiple linear regression, the classification results were weaker, mainly due to the strong class imbalance. Predictions related to the *body score* were particularly poor, with high errors and negative R^2, indicating that the adopted multi-layer perceptrons were not well suited to model this outcome.

The findings of this study also have several implications for social media platforms that host large amounts of health-related content. In areas such as dieting and weight loss, misleading information or harmful narratives may affect particularly vulnerable audiences. The results show that videos with lower levels of informational quality can still achieve high levels of visibility and engagement. This suggests that recommendation systems based

mainly on engagement metrics, such as views or interactions, may not always promote the most reliable or well-supported content.

One possible implication concerns the use of automated tools to support moderation and monitoring processes. The results indicate that large language models can approximate human judgments when evaluating both the quality of information and the level of body-related content. This suggests that such models could be used as scalable tools to help identify potentially problematic videos. Rather than replacing human moderators, automated systems could assist in filtering or prioritizing content that may require closer review, helping moderation teams handle the large volume of uploaded videos more efficiently.

At the same time, the use of large language models may be computationally expensive, especially when applied to very large datasets or in real-time moderation scenarios. For this reason, a preliminary filtering stage could be implemented using lighter analytical approaches. In particular, topic modeling and non-semantic metadata variables may help identify potentially relevant content before applying more computationally demanding models. In this study, these variables showed meaningful correlations with the labels produced by the LLM and demonstrated a moderate predictive capability. Therefore, they could serve as an initial screening layer to reduce the amount of content that needs to be analyzed through more complex language models and human supervision.

Another relevant aspect is the widespread presence of content that is easy to understand but often poorly supported by scientific evidence or external references. Even when such videos do not contain explicit misinformation, they may still contribute to the spread of simplified or anecdotal advice about dieting practices. For this reason, platforms could consider introducing mechanisms that provide users with additional context or disclaimers where not present. For example, videos discussing health topics could be accompanied by links to authoritative sources, informational panels, or references to verified medical information, helping viewers better evaluate the claims presented in the content.

Finally, the dataset revealed the presence of potentially coordinated content patterns, where similar promotional messages appear across multiple channels. In some cases, the same products or claims were promoted by different accounts, suggesting organized promotional activity. This highlights the importance of monitoring not only individual videos but also broader channel behaviors and patterns across the platform.

5.1 Limitations

Despite the insights provided by this study, several limitations should be acknowledged. These limitations concern different stages of the research process, including data collection, annotation procedures, modeling choices, and the generalizability of the obtained results.

A first set of limitations is related to the data collection process. The overall number of collected samples could have been larger in order to improve the robustness and representativeness of the dataset. However, the data collection procedure was constrained by the technical limitations imposed by the transcript retrieval API and by the YouTube Data API v3. In particular, the page token mechanism limits the possibility of exploring

large portions of the platform through automated queries, restricting the amount of retrievable data. As a consequence, the final dataset represents only a partial snapshot of the available content and may not fully capture the variability of videos present on the platform.
In an attempt to partially mitigate this issue, we implemented a retrospective collection procedure to enrich the initial dataset with videos that had already accumulated a higher level of popularity and engagement. This strategy allowed the inclusion of content that users were more likely to encounter and interact with, which can be considered valuable from the perspective of analyzing impactful or widely consumed media. However, this procedure also introduces a potential source of bias. By collecting videos retrospectively, it is possible that some content previously available on the platform had already been removed due to policy violations, copyright issues, other moderation actions, or deletion by the content creator. Consequently, the dataset may underrepresent videos that were deleted after publication, potentially affecting the distribution of certain characteristics and slightly altering the descriptive statistics of the sample.
Another potential source of bias arises from the fact that only videos with available transcripts could be retrieved and analyzed. This constraint was necessary both for the inclusion criteria adopted in the data collection process and for the extraction of textual information from the videos, which enabled the subsequent automated content analysis. However, not all videos on the platform provide transcripts, and the availability of transcripts may depend on factors such as creator behavior, automatic captioning accuracy, or the type of content itself. As a result, videos without transcripts are systematically excluded from the analysis, which may lead to a dataset that is not fully representative of the broader ecosystem of online videos.

Additional limitations concern the use of LLMs for the evaluation of content quality and the level of body-related content. While LLMs offer powerful capabilities for automated text interpretation, their application to these specific evaluative tasks has not yet been extensively validated in the academic literature. In this study, we adopted LLM-assisted annotation as a scalable approach to label a relatively large number of videos. However, the reliability of these labels remains subject to uncertainty, as LLM outputs can be influenced by prompt design, model configuration, and contextual interpretation.
To partially address this potential limitation, we conducted a comparison between LLM-assisted annotations and human annotations on a subset of the dataset. Although this comparison provided useful insights into the alignment between automated and human judgments, the human-labeled sample was limited in size due to time and resource constraints. Moreover, each video in this subset was annotated by only two human labelers. The limited number of annotators restricts the possibility of establishing a robust ground truth, as inter-rater reliability measures are inherently more stable when larger groups of annotators are involved.
Furthermore, the agreement between the two human annotators was on average relatively limited. This outcome suggests that the concepts of content quality and body-related emphasis may be inherently subjective and open to interpretation, particularly considering that the body-related questionnaire adopted in this study has not previously been used in the literature for comparable evaluative purposes. Because of this limited agreement, it becomes difficult to define a fully objective human benchmark against which

automated annotations can be evaluated. Interestingly, the agreement measures between LLM-generated annotations and human annotations were in some cases comparable to, or even higher than, the agreement observed between the two human annotators. While this result might suggest a certain level of consistency in the model outputs, it also raises questions about the extent to which LLM judgments truly approximate human evaluation, or whether they simply reflect systematic patterns introduced by the prompting structure.
Another important limitation related to the use of LLMs concerns the constraints imposed on the prompting strategies and model configurations. Due to limited economic and computational resources, it was not possible to experiment with a wide range of prompting techniques or with more expensive prompting strategies such as few-shot prompting combined with Chain-of-Thought reasoning. These techniques have been shown in previous studies to improve the reasoning capabilities and reliability of LLM outputs in complex evaluation tasks. Therefore, it is possible that alternative prompting configurations could have produced more accurate or more human-aligned annotations.

Limitations also concern the predictive modeling component of the study, particularly the use of multi-layer perceptrons for classification tasks. One of the main issues affecting the performance of the models is the unbalanced distribution of the target classes in the dataset. When one class is significantly more represented than others, machine learning models may struggle to correctly learn patterns associated with the minority class, often resulting in biased predictions toward the majority class. This imbalance likely contributed to the poor predictive performance observed in the experiments.
In addition, the neural network architectures adopted in this study were intentionally simple. The models consisted of basic multi-layer perceptron structures with limited hyperparameter tuning. This design choice was motivated by the intention to explore the predictive capabilities of relatively simple neural architectures and to maintain computational efficiency. However, such simplicity may limit the capacity of the models to capture more complex patterns in the data. More sophisticated architectures or more extensive hyperparameter optimization procedures could potentially lead to improved predictive performance.

The presence of multicollinearity among some of the independent variables included in the multiple linear regression models may also affect the stability and interpretability of the estimated coefficients.

Another limitation concerns the reproducibility of the study. We collected the dataset from a dynamic platform where videos can be removed by creators or by the platform itself at any time. Moreover, the YouTube Data API v3 relies on proprietary algorithms that determine which videos are returned for a given query, meaning that identical API requests performed at different times may produce different results. As a consequence, future attempts to replicate the data collection procedure may lead to a partially different dataset. Reproducibility is also limited by the subjective nature of several stages of the analysis. Human annotations of video quality and body-related content are inherently interpretative, and different annotators may apply evaluation criteria in slightly different ways. Similarly, LLM-based annotations remain probabilistic in nature, even with a relatively low temperature setting. Furthermore, some methodological choices, such as the manual selection of the number of topics based on interpretability considerations, involve a degree of researcher judgment, meaning that different researchers might reasonably obtain

slightly different outcomes when conducting the same analysis.

Finally, the generalizability of the findings should be considered with caution. The dataset analyzed in this study represents a specific subset of videos retrieved through particular search criteria and constrained by the availability of transcripts and metadata. As a result, the observed relationships between video characteristics, topics, and quality scores may not necessarily generalize to the entire ecosystem of online video content or to other platforms beyond YouTube.

Despite these limitations, the methodological approach adopted in this study provides a useful exploratory framework for investigating the relationships between video characteristics, topics, and content quality and level of body centrality. The results obtained should therefore be interpreted as preliminary insights that can inform future research and methodological improvements in the automated analysis of online video content.

5.2 Directions for future research

The results obtained in this study open several interesting directions for future research. Many of these directions naturally emerge from the limitations discussed in the previous section and suggest possible ways to refine both the methodological approach and the analytical scope of the study.

A first important direction concerns the validation of the scores assigned through the assistance of LLMs. Although the comparison with human annotations provided preliminary evidence regarding the reliability of the automated labeling procedure, future research would benefit from a more systematic validation process. In particular, a larger number of human annotators could be involved in the evaluation process, assigning each sampled video to multiple independent labelers. This approach would allow the construction of a more robust ground truth and enable a more precise assessment of the agreement between human judgments and LLM-generated annotations. Moreover, a larger annotated dataset would make it possible to explore in greater depth how different prompting strategies and model configurations affect the reliability of LLM-based evaluations. Future studies could additionally experiment with more advanced prompting techniques, such as few-shot prompting or Chain-of-Thought reasoning, which have been shown to improve the reasoning capabilities of language models in complex evaluation tasks.

Another relevant direction for future research concerns the predictive modeling component of the study. In the present work, relatively simple multi-layer perceptrons were adopted with the specific purpose of exploring the predictive potential of basic neural architectures. While this approach provided useful exploratory insights, future research could investigate whether more sophisticated model architectures, as well as oversampling techniques applied to the minority class to balance the distribution, may lead to improved predictive performance. For instance, deeper neural networks or alternative machine learning models could be tested and compared. More extensive hyperparameter tuning procedures could also be implemented in order to identify model configurations that better capture the underlying patterns in the data. In addition, an interesting extension would involve predicting both the *quality score* and the *body score* simultaneously through multi-output prediction frameworks. Such an approach could capture potential

interactions between the two constructs and provide a more comprehensive representation of the characteristics of video content.
A further improvement in predictive performance could be achieved by expanding the range of input variables used in the models. The results obtained in this study suggest that video topics play an important role in explaining variations in the output scores. Therefore, future research could explore more granular thematic representations, such as subtopics, in order to better characterize the thematic structure of the videos. One possible methodological approach in this direction would be the adoption of topic modeling techniques specifically designed for more fine-grained semantic analysis, such as BERTopic, as suggested by Cheng et al. [2022]. These methods could provide richer representations of video content and potentially improve the predictive capabilities of the models.

The relatively limited associations observed with the *body score* suggest that this dimension may depend more strongly on visual elements of the videos rather than solely on textual content. For this reason, another promising direction for future research would involve incorporating visual information into the analytical framework. Features extracted from video thumbnails, snapshots, or even multiple frames from the video timeline could be used as additional input variables. The integration of textual and visual information would allow for a multimodal analysis of video content, potentially capturing aspects of body-related representation that cannot be inferred from transcripts alone. Advances in computer vision and multimodal machine learning provide several tools that could facilitate this type of analysis.

Another aspect that could be explored in greater depth concerns the role of artificial intelligence in video production. In the present study, we observed a relationship between the use of AI-generated elements and the *quality score*, although the measure of AI usage was based on a small subset of videos. Future research could attempt to develop automated methods for detecting and quantifying the use of AI-generated content in videos. For example, since AI tools are often used to generate background narration or synthetic voice-overs, audio analysis techniques could be employed to identify characteristics associated with synthetic speech. Automated detection techniques could also be developed to identify other forms of AI-generated content, potentially allowing researchers to study the impact of AI-assisted production practices on perceived video quality at a larger scale.

Similarly, future research could further explore the presence of brand mentions or branded products within video content. In this study, the detection of branded content was performed manually on a limited subset of videos. Developing automated methods for identifying product placements, brand mentions, or promotional content could enable more systematic analyses of the relationship between commercial content and video characteristics (Araújo et al. [2017]). In addition, a more detailed categorization of product types or brand categories could provide further insights into how commercial elements interact with video quality, topics, and audience engagement.

Another promising direction concerns the analysis of user engagement. In this study, engagement was measured through view counts and engagement rates derived from standard interaction metrics. However, online platforms provide a much richer set of engagement signals that could be explored in future research. In particular, user comments represent a valuable source of information about audience reactions and perceptions. Textual analysis of comments could reveal how viewers interpret and respond to different types

of content, while network-based analyses could explore the interaction patterns among users within comment sections. Such approaches could provide a deeper understanding of how audiences engage with different types of video content beyond simple quantitative metrics.

Future studies could also explore temporal dynamics in the relationships observed in this research. Online video platforms are characterized by rapidly evolving trends, both in terms of content themes and production practices. Longitudinal analyses could therefore investigate how the relationships between topics, body-related content, and quality evolve over time. For example, researchers could examine whether certain types of content become more or less prevalent over time, or whether audience responses to specific content characteristics change as platform norms and user expectations evolve. A deeper analysis of the data collected through the periodic checks performed during the dynamic data collection process could also provide additional valuable insights.

Beyond these methodological extensions, future research could also explore broader theoretical and analytical perspectives. In particular, the analyses conducted in this study are primarily correlational in nature. While several meaningful relationships between video characteristics and the evaluated scores were identified, future studies could attempt to investigate causal mechanisms more explicitly. For example, experimental or quasi-experimental designs could be employed to better understand whether specific characteristics of video production, presentation style, or thematic focus directly influence perceived content quality or body-related emphasis.

We also simplified the broad concept of risk by considering only two components. Future studies could attempt to define more comprehensive indexes capable of capturing this concept more broadly by incorporating additional relevant factors. Although it is likely impossible to capture and measure all the nuances of risk, analyses including further components could complement the present study and provide additional insights for more accurate content moderation.

Another promising direction concerns the potential role of platform algorithms in shaping the visibility and diffusion of different types of content. Platforms such as YouTube rely on complex recommendation and ranking systems that influence which videos are more likely to be encountered by users. Future research could therefore investigate how algorithmic recommendation mechanisms interact with the characteristics analyzed in this study, potentially amplifying or attenuating the visibility of specific types of content.

Finally, future research could benefit from expanding the scope of the dataset analyzed. The present study focused on a specific subset of videos retrieved through a defined set of queries and constraints. Replicating similar analyses on larger and more diverse datasets would allow researchers to validate the robustness of the findings and explore whether similar patterns emerge across different contexts, for example in different cultural and linguistic scenarios. Online video platforms host content produced for highly diverse audiences, and cultural norms may influence both the production of content and the way it is perceived by viewers. Extending similar analyses to datasets covering different languages, geographic regions, or cultural contexts could provide valuable insights into the extent to which the patterns observed in this study generalize across different segments of the global online video ecosystem. Moreover, extending the analysis to other social media platforms could provide a broader perspective on how video content is produced, characterized, and

evaluated in different online environments. Comparative analyses across platforms could help identify platform-specific dynamics as well as more generalizable patterns in online video production and consumption.

Overall, these directions highlight several opportunities to refine and extend the analytical framework developed in this study. By combining improved annotation procedures, richer data sources, and more advanced modeling techniques, future research could provide a more comprehensive understanding of the factors that shape the characteristics and perceived quality of online video content.

Statement on the use of AI tools

I used generative artificial intelligence tools based on large language models in a limited manner during the preparation of this thesis. Specifically, I employed these tools to assist with language refinement, grammar correction, minor improvements in clarity, and, in a limited number of cases, to summarize passages that I had already written in English.

I did not use AI tools to generate original research ideas, conduct analyses, interpret results, or produce scientific conclusions. All aspects of the research design, data collection (where specified), analysis, and interpretation were carried out by me, with the support and guidance of my supervisors.

The choices regarding the use of these tools were informed by Kosmyna et al. [2025], with the aim of maximizing learning and English writing skills while ensuring the quality of the final result.

Appendix A

Statistical and machine learning tools

In this appendix, we provide a more in-depth theoretical overview of the statistical and machine learning tools employed in the analyses described above and which enabled us to compare different strategies and systematically evaluate the results. We begin by introducing correlation and agreement metrics, then proceed to linear regression and statistical hypothesis testing, and finally discuss neural networks.

A.1 Correlation and agreement metrics

Correlation is a statistical measure that quantifies the strength and direction of the association between two variables. Given two random variables X and Y, correlation evaluates how changes in one variable are related to changes in the other. In general, a positive correlation indicates that larger values of X tend to be associated with larger values of Y, while a negative correlation indicates an inverse relationship. A correlation close to zero suggests little or no systematic association.

In our study, correlation was first employed in addressing the second and third research questions, in order to explore the association between each independent variable and each dependent variable.

Correlation also proved useful when comparing the *quality score* and the *body score* assigned by different human annotators, as well as when comparing human and machine annotations. In this context, the goal is to assess whether higher scores assigned by one annotator tend to correspond to higher scores assigned by another annotator (or by the LLM), and similarly whether lower scores correspond across raters.

For the purposes of this study, Spearman correlation was adopted as the primary metric for measuring agreement. However, correlation is high when two annotators preserve a similar ordering of items, regardless of differences in the absolute values of the scores. When the interest lies in assessing agreement on the exact score values rather than on their relative ranking, alternative agreement measures may be more appropriate, such as Cohen's kappa or the Brennan–Prediger kappa described below.

Pearson correlation coefficient

The Pearson correlation coefficient (ρ for the population, r for a sample) measures the strength and direction of a linear relationship between two variables. In the population case, it is defined as

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}},$$

where $\mathrm{Cov}(X, Y)$ denotes the covariance between X and Y, and $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ are their variances. For a sample of n paired observations (X_i, Y_i), it can be written as

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}},$$

where $\bar{X}$ and $\bar{Y}$ are the sample means. The coefficient ranges from -1 (perfect negative linear relationship) to 1 (perfect positive linear relationship), with 0 indicating no linear association.

Spearman's rank correlation coefficient

Spearman's rank correlation coefficient (ρ_s) is a non-parametric measure of the strength and direction of a *monotonic* relationship between two variables. Given n paired observations (X_i, Y_i), each value is replaced by its rank, denoted $R(X_i)$ and $R(Y_i)$. The coefficient is defined as the Pearson correlation computed on the ranked variables:

$$\rho_s = \frac{\mathrm{Cov}(R(X), R(Y))}{\sqrt{\mathrm{Var}(R(X))}\sqrt{\mathrm{Var}(R(Y))}}.$$

When there are no tied ranks, this expression simplifies to

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

where $d_i = R(X_i) - R(Y_i)$ is the difference between the ranks of the i-th observation.

In the presence of ties (i.e., equal values receiving the same rank), average ranks are assigned to tied observations, and ρ_s is computed directly using the Pearson formula on the ranked data, as in the first equation above. The coefficient ranges from -1 (perfect negative monotonic association) to 1 (perfect positive monotonic association), with 0 indicating no monotonic relationship. Also for Spearman ρ_s denotes the population correlation and r_s its sample estimate.

Cohen's kappa

Cohen's kappa (κ) is a statistic that measures the level of agreement between two annotators who independently classify the same n items into C mutually exclusive categories. Unlike simple percent agreement, κ accounts for the agreement that would be expected by chance. It is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

where p_o is the observed proportion of agreement and p_e is the expected proportion of agreement under chance, computed from the annotators' marginal distributions as follows:

$$p_e = \sum_{c=1}^{C} p_c^{(1)}\, p_c^{(2)},$$

where $p_c^{(1)}$ and $p_c^{(2)}$ denote the marginal proportions of items assigned to category c by annotator 1 and annotator 2, respectively.

The coefficient ranges from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating agreement equivalent to chance.

In the context of comparing single-principle questionnaire scores, the number of categories is $C = 5$, corresponding to the ordered score values. The sixth possible label (`NA`) is not treated as a valid category, as videos for which at least one of the two principle scores is set to `NA` are excluded from the computation.
Given the ordinal structure of the five score categories, a weighted version of Cohen's kappa was adopted in order to account for the degree of disagreement between score values, assigning different penalties depending on how far apart the assigned categories are.

Let $O = (o_{ij})$ be the observed agreement matrix, where o_{ij} denotes the proportion of items that annotator 1 assigns to category i and annotator 2 assigns to category j. Let $E = (e_{ij})$ be the corresponding matrix of expected proportions under chance agreement, computed from the marginal distributions. Introducing a weight matrix $W = (w_{ij})$, which assigns smaller penalties to mild disagreements and larger penalties to more severe ones, the weighted kappa is defined as

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij}\, o_{ij}}{\sum_{i,j} w_{ij}\, e_{ij}}.$$

In our case, we adopted a linear weighting scheme, in which the weights decrease linearly with the distance between categories. Formally, for C ordered categories, linear weights can be defined as

$$w_{ij} = \frac{|i - j|}{C - 1},$$

so that disagreements between adjacent categories receive a smaller penalty, while disagreements between categories that are farther apart are penalized proportionally more.

The interpretation of κ values follows the conventional benchmarks proposed by Landis and Koch [1977]. Although originally introduced for the unweighted Cohen's kappa, these thresholds are commonly adopted also for weighted versions of the statistic. According to this scale:

- $\kappa < 0.00$: Poor agreement
- $\kappa = 0.00$: No agreement
- $0.01 \leq \kappa \leq 0.20$: Slight agreement
- $0.21 \leq \kappa \leq 0.40$: Fair agreement

- $0.41 \leq \kappa \leq 0.60$: Moderate agreement
- $0.61 \leq \kappa \leq 0.80$: Substantial agreement
- $0.81 \leq \kappa \leq 1.00$: Almost perfect agreement

It should be noted, however, that these benchmarks are heuristic guidelines rather than theoretically grounded cut-offs.

Brennan-Prediger kappa

The Brennan–Prediger kappa (κ_{BP}, Brennan and Prediger [1981]) is an inter-rater agreement coefficient proposed as an alternative to Cohen's kappa, particularly in situations where Cohen's kappa may be affected by prevalence or marginal distribution imbalances. Unlike Cohen's kappa, which computes the expected agreement from the observed marginal proportions, the Brennan–Prediger coefficient assumes that all categories are equally likely under chance agreement.

The expected agreement under chance is defined as $p_e = \frac{1}{C}$. The Brennan–Prediger kappa is then computed as

$$\kappa_{BP} = \frac{p_o - \frac{1}{C}}{1 - \frac{1}{C}}.$$

By fixing the expected agreement to $1/C$, κ_{BP} reduces the influence of skewed marginal distributions and often provides more stable agreement estimates when category prevalence is highly unbalanced. However, it is generally defined for nominal categories and does not naturally incorporate a weighted version for ordinal data.

Based on previous literature (Khalil et al. [2025]), the interpretation benchmarks described above can also be applied, with appropriate caution, to κ_{BP}.

Mean absolute error for ordinal agreement

When the categories are ordinal, inter-rater agreement can also be evaluated by measuring the average magnitude of disagreement between annotators. Let X_i and Y_i denote the scores assigned by two annotators to the i-th item, with $i = 1, \ldots, n$. The Mean Absolute Error (MAE) is defined as

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |X_i - Y_i| .$$

MAE measures the average absolute distance between the two sets of ratings. A value of 0 indicates perfect agreement, while larger values reflect greater average disagreement. Unlike chance-corrected coefficients such as Cohen's kappa, MAE does not adjust for expected agreement under chance; instead, it directly quantifies the typical severity of disagreement, making it particularly suitable when the magnitude of ordinal discrepancies is of primary interest.

A.2 Multiple linear regression

Multiple linear regression is a statistical method used to model the relationship between one numerical dependent variable and multiple independent variables. The goal is to explain or predict the value of a response variable Y as a linear combination of p predictors $X_1, \ldots, X_p$.

Formally, the model can be written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i,$$

where Y_i is the outcome for observation i, β_0 is the intercept, $\beta_1, \ldots, \beta_p$ are the regression coefficients, and ε_i is the error term. The coefficients represent the expected change in Y associated with a one-unit increase in the corresponding predictor, holding all other variables constant.

The parameters are typically estimated, as in our case, using the ordinary least squares (OLS) method, which minimizes the sum of squared residuals, i.e., the differences between observed and predicted values. Multiple linear regression allows for the simultaneous assessment of several predictors, making it possible to evaluate their individual contributions while controlling for the others.

To have an overall indication of the model's goodness of fit we used the coefficient of determination, denoted by R^2. It measures the proportion of variance in the dependent variable that is explained by the regression model.
Formally, R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2},$$

where Y_i are the observed values, $\hat{Y}_i$ are the predicted values from the model, and $\bar{Y}$ is the sample mean of the dependent variable. The numerator represents the residual sum of squares (unexplained variability), while the denominator represents the total sum of squares (total variability).
The value of R^2 generally ranges between 0 and 1, with higher values indicating that a larger proportion of the variability in the response variable is accounted for by the predictors included in the model.

However, when multiple predictors are included, R^2 tends to increase as additional variables are added to the model, even if those variables do not provide meaningful explanatory power. For this reason, it is common to consider the adjusted coefficient of determination, denoted as R^2_{adj}. This metric penalizes the inclusion of unnecessary predictors by accounting for the number of explanatory variables relative to the sample size.
Formally, it is defined as

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}.$$

Unlike R^2, the adjusted R^2 may decrease when a new variable is added if that variable does not improve the model sufficiently. Therefore, R^2_{adj} provides a more reliable measure of goodness of fit when comparing models with different numbers of predictors.

An important issue in multiple linear regression is multicollinearity, which occurs when two or more predictors are highly linearly correlated. In the presence of strong collinearity,

the estimation of regression coefficients may become unstable. As a result, it becomes difficult to isolate the individual contribution of correlated predictors, even if the overall model fit remains acceptable.

Variance inflation factor and iterative variable removal

A common diagnostic measure for multicollinearity is the Variance Inflation Factor (VIF). For the j-th predictor, the VIF is defined as

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination obtained by regressing the j-th predictor on all the remaining predictors. The VIF quantifies how much the variance of the estimated coefficient $\hat{\beta}_j$ is inflated due to collinearity. Values close to 1 indicate negligible collinearity, while larger values (commonly above 5 or 10) suggest problematic levels of multicollinearity.

To address multicollinearity, we used the practical strategy of iteratively removing variables based on the VIF. In this procedure, VIF values are first computed for all predictors; the variable with the highest VIF exceeding a predefined threshold of VIF=10 is removed. The VIFs are then recomputed on the reduced set of predictors, and the process is repeated until all remaining variables exhibit acceptable VIF values. This approach helps obtain a more stable and interpretable regression model.

Categorical variables and dummy encoding

In multiple linear regression, categorical predictors cannot be included directly in their original form, as the model requires numerical inputs. To incorporate a categorical variable with K distinct categories, dummy encoding is applied.

Under dummy encoding, the categorical variable is transformed into $K - 1$ binary variables. Each dummy variable takes value 1 if the observation belongs to a given category and 0 otherwise. One category is omitted and serves as the reference (or baseline) category, preventing perfect multicollinearity.

Each coefficient β_j of a dummy variable D_{ij} represents the expected difference in the dependent variable between category j and the reference category, holding all other predictors constant.

A.3 Statistical tests

A statistical test is a formal procedure used to assess evidence about a hypothesis on the basis of observed data. Typically, one specifies a null hypothesis (H_0), representing a baseline assumption (e.g., no effect or no association), and an alternative hypothesis (H_1), representing a competing claim. The test statistic is computed from the sample data and compared against its sampling distribution under the assumption that H_0 is true, in order to evaluate how compatible the observed data are with the null hypothesis.

The p-value quantifies this compatibility. Formally, it is the probability, under the assumption that H_0 holds, of observing a result at least as extreme as the one obtained. Small p-values indicate that the observed data would be unlikely if the null hypothesis were true, and therefore provide evidence against H_0. However, the p-value does not measure the probability that H_0 is true, nor does it quantify the magnitude or practical importance of an effect.

The significance level, denoted by α, represents the pre-specified threshold used to decide whether a result is considered statistically significant. It corresponds to the maximum probability of committing a Type I error, that is, rejecting the null hypothesis when it is in fact true. In our analyses, the significance level was set to the conventional value of $\alpha = 0.05$.

Bonferroni correction

When multiple statistical tests are performed simultaneously, the probability of incurring at least one Type I error increases with the number of tests conducted. If m tests are carried out at significance level α, the probability of obtaining at least one statistically significant result purely by chance exceeds α.

The Bonferroni correction is a conservative procedure designed to control the family-wise error rate, that is, the probability of making at least one Type I error across all tests. The correction can be implemented by adjusting the p-value obtained from each test, multiplying it by the total number of tests performed:

$$p_{\text{adj}} = m \cdot p.$$

A result is considered statistically significant if $p_{\text{adj}} < \alpha$. Equivalently, the correction can be applied by testing each hypothesis at a reduced significance level α/m.

Statistical test for Spearman's correlation

The statistical test for Spearman's rank correlation coefficient evaluates whether the observed monotonic association between two variables differs significantly from zero.

The null and alternative hypotheses are defined as

$$H_0 : \rho_s = 0 \quad \text{and} \quad H_1 : \rho_s \neq 0.$$

In other contexts, the alternative hypothesis may also be one-sided (i.e., $\rho_s < 0$ or $\rho_s > 0$), although this is not considered in the present analysis.

For moderate to large sample sizes, the test statistic is computed as

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \;\sim\; t_{n-2} \quad \text{under } H_0,$$

meaning that, under the null hypothesis, it approximately follows a Student's t distribution with $n-2$ degrees of freedom. The corresponding p-value is obtained from this reference distribution.

Statistical test for multiple linear regression coefficients

In multiple linear regression, statistical tests are used to assess whether each predictor has a significant association with the dependent variable, controlling for the other variables in the model.

For each regression coefficient β_j, the null and alternative hypotheses are defined as

$$H_0 : \beta_j = 0 \qquad \text{and} \qquad H_1 : \beta_j \neq 0.$$

Also for this type of test, the alternative hypothesis may be formulated as one-sided.

Let $\hat{\beta}_j$ denote the estimated coefficient and $\mathrm{SE}(\hat{\beta}_j)$ its standard error

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 \, (X^\top X)^{-1}_{jj}}, \qquad \hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

The test statistic is computed as

$$t = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \;\sim\; t_{n-p-1} \quad \text{under } H_0,$$

where n is the number of observations and p is the number of predictors included in the model (excluding the intercept).

Mann–Whitney U test

To compare the distributions of two independent groups the non-parametric Mann–Whitney U test has been used. Let $X_1, \ldots, X_{n_1}$ be an i.i.d. sample drawn from the first group and $Y_1, \ldots, Y_{n_2}$ an i.i.d. sample drawn from the second group, than the test hypotheses are:

- H_0: the two groups have the same distribution;
- H_1: the two groups differ in terms of position.

The test does not require normality of distributions.

If all $n_1 + n_2$ observations are jointly ranked in increasing order and R_1 and R_2 denote the sums of ranks for the first and second group, respectively, the test statistics are defined as

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}, \qquad U_2 = R_2 - \frac{n_2(n_2+1)}{2}.$$

These quantities satisfy $U_1 + U_2 = n_1 n_2$, and therefore it is sufficient to compute only one of them. The statistic U can be interpreted as the number of pairwise comparisons (X_i, Y_j) in which the observation from the first group precedes (i.e., is smaller than) the observation from the second group. It follows that $0 \leq U \leq n_1 n_2$.
For sufficiently large sample sizes, the standardized statistic

$$Z = \frac{U - \mu_U}{\sigma_U} \qquad \text{with} \qquad \mu_U = \frac{n_1 n_2}{2}, \qquad \sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

is approximately normally distributed under H_0,

$$Z \sim \mathcal{N}(0,1).$$

In the case of a two-sided test, as adopted in this study, the p-value is computed as

$$p = 2\,\mathbb{P}(|Z| \geq |z_{\text{obs}}|),$$

which depends only on the absolute value of Z. Since $U_2 = n_1 n_2 - U_1$, it follows that the corresponding standardized statistics satisfy $Z_2 = -Z_1$, and therefore the two-sided p-value is identical whether U_1 or U_2 is used.

An effect size measure can be obtained as

$$r = \frac{Z}{\sqrt{n}},$$

where $n = n_1 + n_2$ is the total sample size. According to Cohen's conventional benchmarks, $|r| \approx 0.1$ indicates a small effect, $|r| \approx 0.3$ a medium effect, and $|r| \approx 0.5$ a large effect.

The Mann–Whitney statistic is also directly related to the Area Under the ROC Curve (AUC). Specifically,

$$\text{AUC} = \frac{U}{n_1 n_2},$$

which corresponds to the probability that a randomly selected observation from one group has a higher value than a randomly selected observation from the other group.

Kruskal–Wallis H test

The Kruskal–Wallis test is a non-parametric method used to assess whether $k \geq 2$ independent groups originate from the same distribution. It represents a rank-based extension of the Mann–Whitney test to more than two groups.

Let n_i denote the sample size of the i-th group, with total sample size $n = \sum_{i=1}^{k} n_i$. All n observations are jointly ranked in increasing order. Let R_i be the sum of ranks for the i-th group. The test statistic is defined as

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1).$$

Under the null hypothesis that all groups share the same distribution, and for sufficiently large sample sizes, the statistic approximately follows a chi-square distribution with $k-1$ degrees of freedom:

$$H \sim \chi^2_{k-1}.$$

The corresponding p-value is obtained from this reference distribution. However, the test does not specify which groups differ, and post-hoc comparisons are required to identify pairwise differences.

A.4 Multi-layer perceptron

To address the predictive task in our study we used Multi-Layer Perceptrons (MLPs), feedforward artificial neural networks composed of multiple layers of interconnected neurons. It extends the single-layer perceptron by introducing one or more hidden layers, allowing the model to learn complex, non-linear relationships between input features and outputs.

An MLP typically consists of an input layer, one or more hidden layers, and an output layer. Let $\mathbf{x} \in \mathbb{R}^q$ denote the input vector. Each neuron in a hidden layer computes a weighted linear combination of its inputs, followed by a non-linear activation function:

$$\mathbf{h}^{(\ell)} = \sigma\left(W^{(\ell)}\mathbf{h}^{(\ell-1)} + \mathbf{b}^{(\ell)}\right),$$

where $W^{(\ell)}$ and $\mathbf{b}^{(\ell)}$ are the weight matrix and bias vector of layer ℓ, respectively, $\sigma(\cdot)$ is a non-linear activation function, and $\mathbf{h}^{(0)} = \mathbf{x}$.

The final output layer produces the prediction:

$$\hat{\mathbf{y}} = f\left(W^{(L)}\mathbf{h}^{(L-1)} + \mathbf{b}^{(L)}\right),$$

where L denotes the total number of layers and $f(\cdot)$ depends on the task.

Model parameters are estimated by minimizing a loss function using gradient-based optimization methods such as stochastic gradient descent. Gradients are computed efficiently through the backpropagation algorithm.

Due to the presence of non-linear activation functions and multiple hidden layers, MLPs are universal function approximators, meaning that, under mild conditions, they can approximate any continuous function on a compact domain. However, their performance depends on appropriate architecture design and sufficient training data.

MLPs can be applied to both classification and regression tasks depending on the form of the output layer and the loss function used during training. In classification problems, the model predicts discrete class labels or class probabilities, typically using a sigmoid activation for binary classification or a softmax function for multiclass settings. Model performance is commonly evaluated using metrics derived from the confusion matrix. Let TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. The following metrics are used:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN},$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \qquad \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

In regression tasks, the output layer produces a continuous prediction, typically using an identity activation function. Model performance is evaluated using error-based metrics that measure the discrepancy between predicted and observed values. Let y_i denote the true value and $\hat{y}_i$ the predicted value for observation i, with n observations. We consider the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination R^2:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \qquad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where $\bar{y}$ denotes the sample mean of the observed values.

Appendix B

Correlation scatter plots

This appendix visually explores the relationships between the two main dependent variables, *body score* and *quality score*, and their strongest numerical predictors.

Figure B.1 and Figure B.2 display associations with non-semantic variables, log-transforming highly skewed data to improve visualization. Figure B.3 and Figure B.4 show the relationships with NMF topic-specific scores. In these plots, orange points highlight videos formally assigned to the respective topic, while blue points represent all other samples.

Figure B.1: Scatter plots of most correlated non-semantic variables scores versus *body score*. The solid black line indicates the fitted OLS regression line.

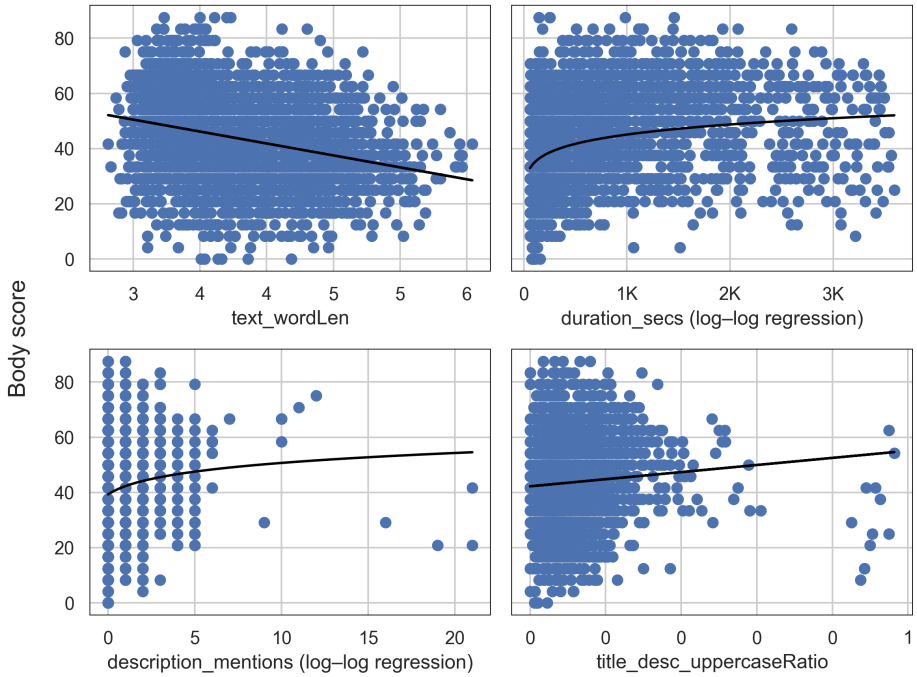


Figure B.2: Scatter plots of most correlated non-semantic variables scores versus *quality score*. The solid black line indicates the fitted OLS regression line.

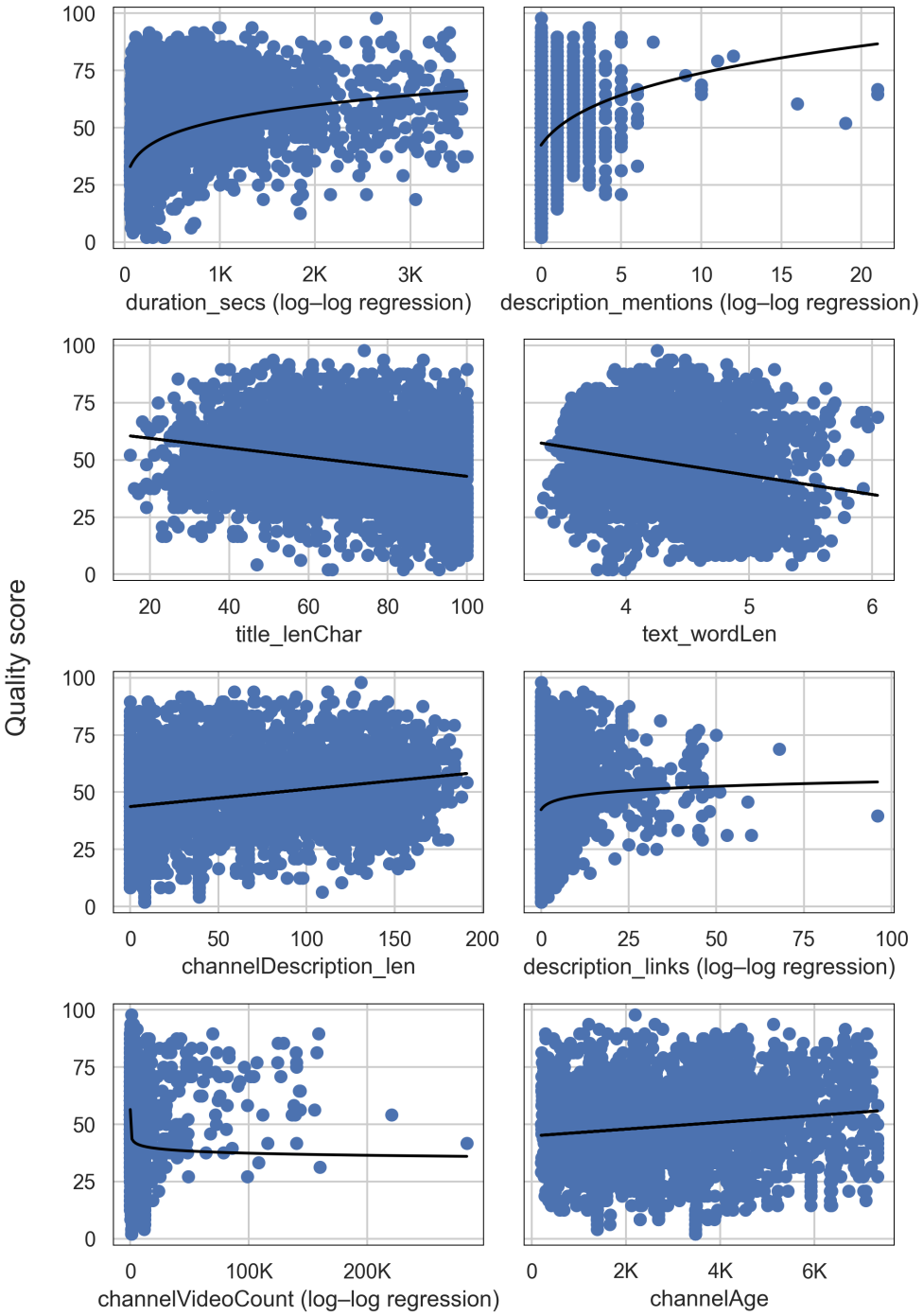


Figure B.3: Scatter plots of most correlated topic-specific scores versus *body score*. Orange points represent samples associated with the corresponding topic, while blue points denote all other samples. The solid black line indicates the fitted OLS regression line.

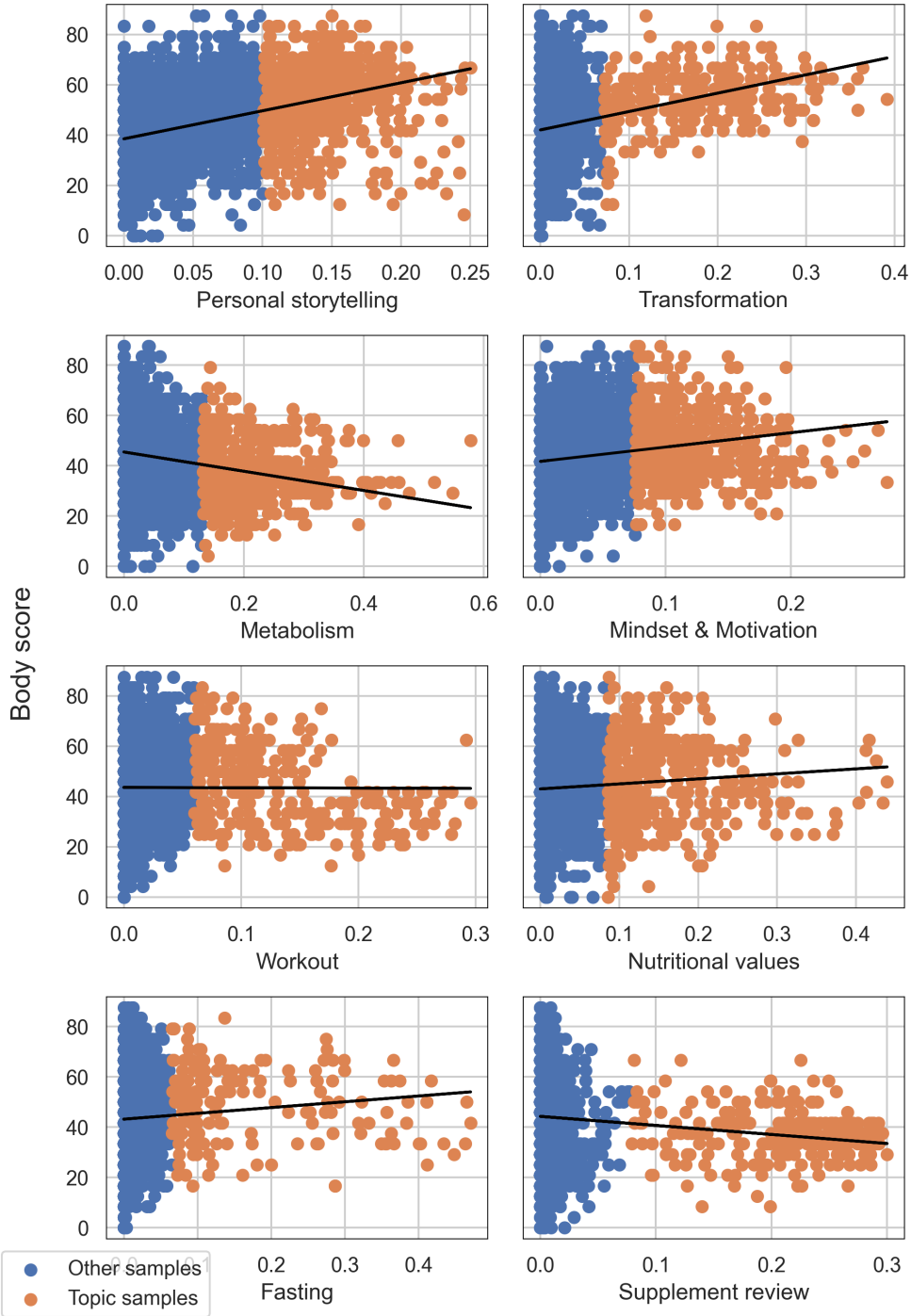
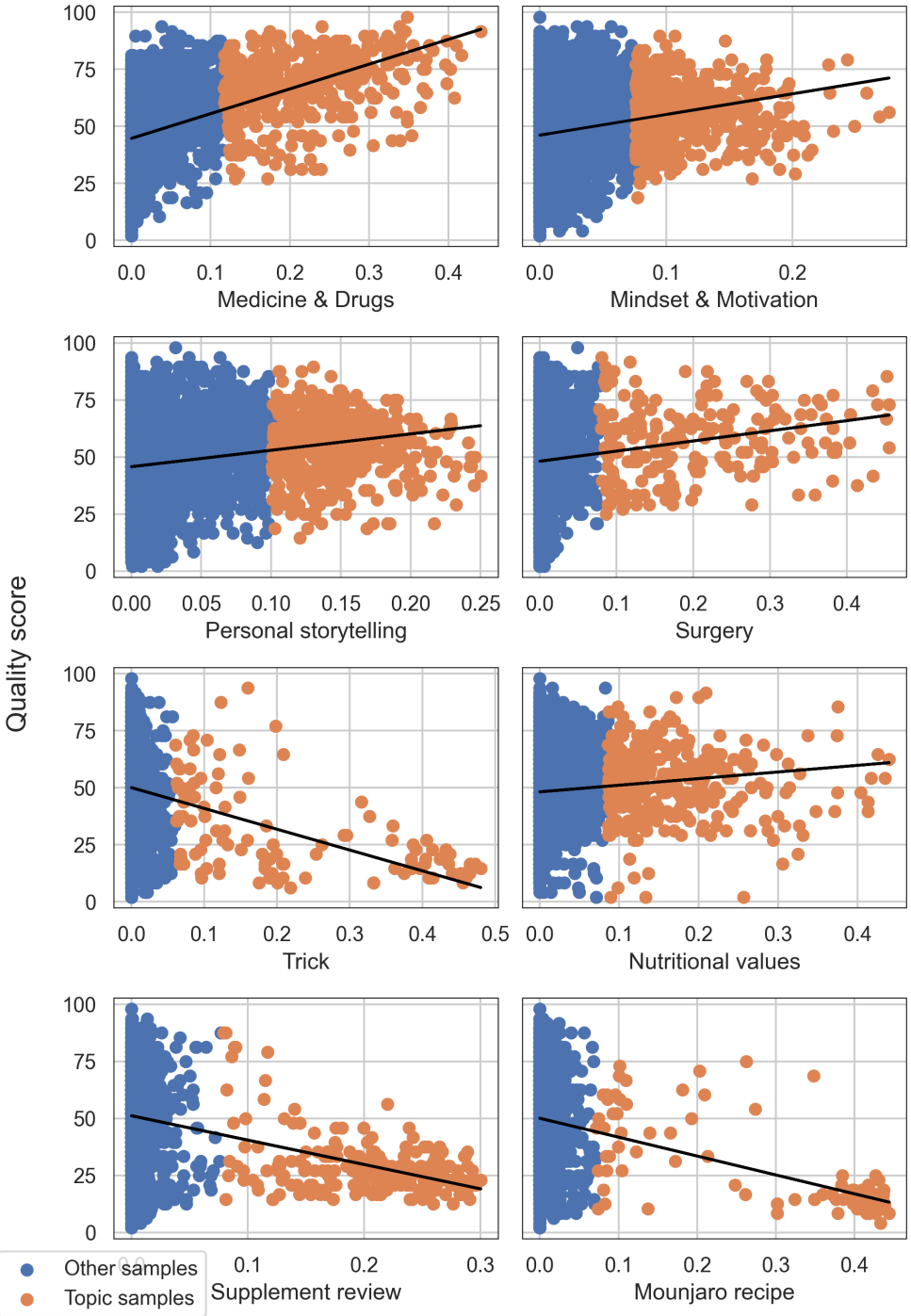


Figure B.4: Scatter plots of most correlated topic-specific scores versus *quality score*. Orange points represent samples associated with the corresponding topic, while blue points denote all other samples. The solid black line indicates the fitted OLS regression line.



Bibliography

Youtube API Reference. https://developers.google.com/youtube/v3/docs?hl=en. Accessed: 2025-11-25.

Abeer S Alzaben, Khawlah I Alzaidy, Mona A Alghamdi, Raghad A Alanzi, Rawan T Aljohari, Reema A Alahaideb, and Nada Benajiba. The use of social media to search for weight reduction information: Assessment of the perception among a sample of Saudi adults. *Digital Health*, 8:20552076221136939, November 2022. ISSN 2055-2076. doi: 10.1177/20552076221136939. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9638696/.

Monica Anderson, Michelle Faverio, and Jeffrey Gottfried. Teens, Social Media and Technology 2023: YouTube, TikTok, Snapchat and Instagram remain the most widely used online platforms among U.S. teens. Technical report, Pew Research Center, 2023. URL https://www.jstor.org/stable/resrep63510.

Camila Araújo, Gabriel Magno, Wagner Meira Jr, Virgilio Almeida, Pedro Hartung, and Danilo Doneda. Characterizing videos, audience and advertising in Youtube channels for kids. July 2017. doi: 10.48550/arXiv.1707.00971.

Heger Arfaoui, Mohammed Iheb Hergli, Beya Benzina, and Slimane BenMiled. A Reproducible Framework for Neural Topic Modeling in Focus Group Analysis, November 2025. URL https://arxiv.org/abs/2511.18843v2.

C. H. Basch, I. C. H. Fung, A. Menafro, C. Mo, and J. Yin. An exploratory assessment of weight loss videos on YouTube™. *Public Health*, 151:31–38, October 2017. ISSN 0033-3506. doi: 10.1016/j.puhe.2017.06.016. URL https://www.sciencedirect.com/science/article/pii/S0033350617302196.

Nazlı Batar, Seda Kermen, Sezen Sevdin, Nida Yıldız, and Duygu Güçlü. Assessment of the Quality and Reliability of Information on Nutrition After Bariatric Surgery on YouTube. *Obesity Surgery*, 30(12):4905–4910, December 2020. ISSN 1708-0428. doi: 10.1007/s11695-020-05015-z. URL https://doi.org/10.1007/s11695-020-05015-z.

Nada Benajiba, Maha Alhomidi, Fahdah Alsunaid, Aljawharah Alabdulkarim, Elizabeth Dodge, Enmanuel A. Chavarria, and Basil H. Aboul-Enein. Video clips of the Mediterranean Diet on YouTube TM: A social Media Content Analysis. *American Journal of Health Promotion*, 37(3):366–374, March 2023. ISSN 0890-1171. doi:

10.1177/08901171221132113. URL https://doi.org/10.1177/08901171221132113. Publisher: SAGE Publications Inc.

Robert L Brennan and Dale J Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699, 1981.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

Michelle I. Cardel, Sarah Chavez, Jiang Bian, Eribeth Peñaranda, Darci R. Miller, Tianyao Huo, and François Modave. Accuracy of weight loss information in Spanish search engine results on the internet. *Obesity*, 24(11):2422–2434, 2016. ISSN 1930-739X. doi: 10.1002/oby.21646. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/oby.21646. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/oby.21646.

Kathy Charmaz. *Constructing Grounded Theory*.

Deborah Charnock. *The DISCERN handbook: quality criteria for consumer health information on treatment choices*. Radcliffe Medical, Abingdon, 1998. ISBN 978-1-85775-310-3. OCLC: 59581385.

Keyi Cheng, Stefan Inzer, Adrian Leung, Xiaoxian Shen, Michael Perlmutter, Michael Lindstrom, Joyce Chew, Todd Presner, and Deanna Needell. Multi-scale Hybridized Topic Modeling: A Pipeline for Analyzing Unstructured Text Datasets via Topic Modeling, November 2022. URL https://arxiv.org/abs/2211.13496v1.

Minh Duc Chu, Zihao He, Rebecca Dorn, and Kristina Lerman. Large Language Models Help Reveal Unhealthy Diet and Body Concerns in Online Eating Disorders Communities, May 2024. URL http://arxiv.org/abs/2401.09647. arXiv:2401.09647 [cs].

Wu Chunqiong, Jiang Shan, Sun Jianhong, and Liu Yingqi. Impact of YouTube User-Generated Content on News Dissemination and Youth Information Reception. *Health Expectations : An International Journal of Public Participation in Health Care and Health Policy*, 28(5):e70408, September 2025. ISSN 1369-6513. doi: 10.1111/hex.70408. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12399985/.

Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. Falling into the echo chamber: the italian vaccination debate on twitter. In *Proceedings of the International AAAI conference on web and social media*, volume 14, pages 130–140, 2020.

Ligia Alfaro Cruz, Isha Kaul, Yan Zhang, Robert Jay Shulman, and Bruno Pedro Chumpitazi. ASSESSMENT OF QUALITY AND READABILITY OF INTERNET DIETARY INFORMATION ON IRRITABLE BOWEL SYNDROME. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*, 17(3):566–567, February 2019. ISSN 1542-3565. doi: 10.1016/j.cgh.2018.05.018. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6250590/.

Camilla Lindvall Dahlgren, Christine Sundgot-Borgen, Ingela Lundin Kvalem, Anne-Louise Wennersberg, and Line Wisting. Further evidence of the association between social media use, eating disorder pathology and appearance ideals and pressure: a cross-sectional study in norwegian adolescents. *Journal of Eating Disorders*, 12(1):34, 2024.

Xinran Dai and Jing Wang. Effect of online video infotainment on audience attention. *Humanities and Social Sciences Communications*, 10(1):421, July 2023. ISSN 2662-9992. doi: 10.1057/s41599-023-01921-6. URL https://www.nature.com/articles/s41599-023-01921-6. Publisher: Palgrave.

Alexandra Dane and Komal Bhatia. The social media diet: A scoping review to investigate the association between social media, body image and eating disorders amongst young people. *PLOS Global Public Health*, 3(3):e0001091, March 2023. ISSN 2767-3375. doi: 10.1371/journal.pgph.0001091. URL https://journals.plos.org/globalpublichealth/article?id=10.1371/journal.pgph.0001091. Publisher: Public Library of Science.

Caitlin Davey, Emily Newman, Joanna Hare, David Fluck, and Thang Sieu Han. Risk of instagram dieting trends on eating behaviour and body satisfaction in women of different age and body mass index. *Journal of Technology in Behavioral Science*, pages 1–10, 2024.

Emily Denniss, Rebecca Lindberg, and Sarah A. McNaughton. Development of Principles for Health-Related Information on Social Media: Delphi Study. *Journal of Medical Internet Research*, 24(9):e37337, September 2022. doi: 10.2196/37337. URL https://www.jmir.org/2022/9/e37337. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Emily Denniss, Rebecca Lindberg, Laura E. Marchese, and Sarah A. McNaughton. #Fail: the quality and accuracy of nutrition-related information by influential Australian Instagram accounts. *The International Journal of Behavioral Nutrition and Physical Activity*, 21(1):16, February 2024. ISSN 1479-5868. doi: 10.1186/s12966-024-01565-y.

Digital News Report. Digital News Report 2023 | Reuters Institute for the Study of Journalism, June 2023. URL https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023.

Philippine Fassier, Anne-Sophie Chhim, Valentina A Andreeva, Serge Hercberg, Paule Latino-Martel, Camille Pouchieu, and Mathilde Touvier. Seeking health-and nutrition-related information on the internet in a large population of french adults: results of the nutrinet-santé study. *British Journal of Nutrition*, 115(11):2039–2046, 2016.

Bruce Fraser. An approach to discourse markers. *Journal of Pragmatics*, 14(3):383–398, June 1990. ISSN 0378-2166. doi: 10.1016/0378-2166(90)90096-V. URL https://www.sciencedirect.com/science/article/pii/037821669090096V.

K Gkouskou, A Markaki, M Vasilaki, A Roidis, and I Vlastos. Quality of nutritional information on the Internet in health and disease. *Hippokratia*, 15(4):304–307, 2011. ISSN 1108-4189. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3876843/.

Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, March 2022. URL http://arxiv.org/abs/2203.05794. arXiv:2203.05794 [cs].

Rocío Guardiola-Wanden-Berghe, Josefa D. Gil-Pérez, Javier Sanz-Valero, and Carmina Wanden-Berghe. Evaluating the quality of websites relating to diet and eating disorders. *Health Information & Libraries Journal*, 28 (4):294–301, 2011. ISSN 1471-1842. doi: 10.1111/j.1471-1842.2011.00961.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1471-1842.2011.00961.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1471-1842.2011.00961.x.

Reiko Hirasawa, Kazumi Saito, Yoko Yachi, Yoko Ibe, Satoru Kodama, Mihoko Asumi, Chika Horikawa, Aki Saito, Yoriko Heianza, Kazuo Kondo, Hitoshi Shimano, and Hirohito Sone. Quality of Internet information related to the Mediterranean diet. *Public Health Nutrition*, 15(5):885–893, May 2012. ISSN 1475-2727, 1368-9800. doi: 10.1017/S1368980011002345. URL https://www.cambridge.org/core/journals/public-health-nutrition/article/quality-of-internet-information-related-to-the-mediterranean-diet/C62DD21FF6E822D1A625064A78E11EAE.

Grace Holland and Marika Tiggemann. A systematic review of the impact of the use of social networking sites on body image and disordered eating outcomes. *Body Image*, 17:100–110, June 2016. ISSN 1740-1445. doi: 10.1016/j.bodyim.2016.02.008. URL https://www.sciencedirect.com/science/article/pii/S1740144516300912.

Colin Horning. Social Media News Consumption by College Students Using the Elaboration Likelihood Model. *Dissertations and Theses @ UNI*, January 2024. URL https://scholarworks.uni.edu/etd/1702.

Flavio Jeronimo and Eliana Veiga Carraca. Effects of fitspiration content on body image: a systematic review. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, 27(8):3017–3035, 2022.

Paweł Kabata, Dorota Winniczuk-Kabata, Piotr Maciej Kabata, Janusz Jaśkiewicz, and Karol Połom. Can Social Media Profiles Be a Reliable Source of Information on Nutrition and Dietetics? *Healthcare*, 10(2):397, February 2022. ISSN 2227-9032. doi: 10.3390/healthcare10020397. URL https://www.mdpi.com/2227-9032/10/2/397. Publisher: Multidisciplinary Digital Publishing Institute.

Katikapalli Subramanyam Kalyan. A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4, October 2023. URL https://papers.ssrn.com/abstract=4593895.

Mahmoud Khalil, Fatma Mohamed, and Abdulhadi Shoufan. Evaluating the quality of medical content on YouTube using large language models. *Scientific Reports*, 15(1):9906, March 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-94208-6. URL https://www.nature.com/articles/s41598-025-94208-6.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4, 2025.

Megan Kreft, Brittany Smith, Daniella Hopwood, and Renee Blaauw. The use of social media as a source of nutrition information. *South African Journal of Clinical Nutrition*, 36(4):162–168, December 2023. doi: 10.1080/16070658.2023.2175518. URL https://journals.co.za/doi/full/10.1080/16070658.2023.2175518. Publisher: NISC (Pty) Ltd.

Leonardo La Rocca. Multimodal LLMs vs LLMs vs RoBERTa: evaluating AI performance in detecting conspiracies on YouTube. April 2025. URL https://www.politesi.polimi.it/handle/10589/234218. Accepted: 2025-07-02T09:51:10Z.

Viet Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Nguyen. BehanceCC: A ChitChat Detection Dataset For Livestreaming Video Transcripts. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7284–7290, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.791/.

Christopher Lalk, Tobias Steinbrenner, Weronika Kania, Alexander Popko, Robin Wester, Jana Schaffrath, Steffen Eberhardt, Brian Schwartz, Wolfgang Lutz, and Julian Rubel. Measuring Alliance and Symptom Severity in Psychotherapy Transcripts Using Bert

Topic Modeling. *Administration and Policy in Mental Health and Mental Health Services Research*, 51(4):509–524, July 2024. ISSN 1573-3289. doi: 10.1007/s10488-024-01356-4. URL https://doi.org/10.1007/s10488-024-01356-4.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

Joanne Mayoh and Ian Jones. Young people's experiences of engaging with fitspiration on Instagram: Gendered perspective. *Journal of Medical Internet Research*, 23(10), 2021. ISSN 1438-8871. doi: 10.2196/17811. Place: Canada Publisher: JMIR Publications.

Yelena Mejova and Víctor Suarez-Lledó. Impact of online health awareness campaign: Case of national eating disorders association. In *International Conference on Social Informatics*, pages 192–205. Springer, 2020.

Yelena Mejova and Michele Tizzani. Vaccine Hesitancy on YouTube: a Competition between Health and Politics, July 2025. URL https://ui.adsabs.harvard.edu/abs/2025arXiv250707517M. ADS Bibcode: 2025arXiv250707517M.

Marisa Minadeo and Lizzy Pope. Weight-normative messaging predominates on TikTok—A qualitative content analysis. *PLOS ONE*, 17(11):e0267997, November 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0267997. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0267997. Publisher: Public Library of Science.

Beki Moult, Linda S Franck, and Helen Brady. Ensuring Quality Information for Patients: development and preliminary validation of a new instrument to improve the quality of written health care information. *Health Expectations*, 7(2):165–175, 2004. ISSN 1369-7625. doi: 10.1111/j.1369-7625.2004.00273.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1369-7625.2004.00273.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1369-7625.2004.00273.x.

Emily Munro, Gabriella Wells, Rigel Paciente, Nicole Wickens, Daniel Ta, Joelie Mandzufas, Karen Lombardi, and Alix Woolard. Diet culture on TikTok: a descriptive content analysis. *Public Health Nutrition*, 27(1):e169, January 2024. ISSN 1368-9800, 1475-2727. doi: 10.1017/S1368980024001381. URL https://www.cambridge.org/core/journals/public-health-nutrition/article/diet-culture-on-tiktok-a-descriptive-content-analysis/B8B5F4843393D5702EAA3B8C75603AE0.

Faisal A Nawaz, Mehr Muhammad Adeel Riaz, Nimrat ul ain Banday, Aakanksha Singh, Zara Arshad, Hanan Derby, and Meshal A Sultan. Social media use among adolescents with eating disorders: a double-edged sword. *Frontiers in psychiatry*, 15:1300182, 2024.

NIMH. Eating Disorders - National Institute of Mental Health (NIMH). URL https://www.nimh.nih.gov/health/topics/eating-disorders.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on Medical Challenge Problems, April 2023. URL http://arxiv.org/abs/2303.13375. arXiv:2303.13375 [cs].

A. Ostry, M. L. Young, and M. Hughes. The quality of nutritional information available on popular websites: a content analysis. *Health Education Research*, 23(4):648–655, September 2007. ISSN 0268-1153, 1465-3648. doi: 10.1093/her/cym050. URL https://academic.oup.com/her/article-lookup/doi/10.1093/her/cym050.

Soham Parikh, Quaizar Vohra, Prashil Tumbade, and Mitul Tiwari. Exploring Zero and Few-shot Techniques for Intent Classification, May 2023. URL http://arxiv.org/abs/2305.07157. arXiv:2305.07157 [cs].

Rebecca L. Pearl and Rebecca M. Puhl. The distinct effects of internalizing weight bias: An experimental study. *Body Image*, 17:38–42, June 2016. ISSN 1740-1445. doi: 10.1016/j.bodyim.2016.02.002. URL https://www.sciencedirect.com/science/article/pii/S1740144515300085.

Raquel Franzini Pereira and Marle Alvarenga. Disordered eating: identifying, treating, preventing, and differentiating it from eating disorders. *Diabetes Spectrum*, 20(3):141–148, 2007.

Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA, February 2015. Association for Computing Machinery. ISBN 978-1-4503-3317-7. doi: 10.1145/2684822.2685324. URL https://dl.acm.org/doi/10.1145/2684822.2685324.

Alyssa N. Saiphoo and Zahra Vahedi. A meta-analytic review of the relationship between social media use and body image disturbance. *Computers in Human Behavior*, 101:259–275, December 2019. ISSN 0747-5632. doi: 10.1016/j.chb.2019.07.028. URL https://www.sciencedirect.com/science/article/pii/S0747563219302717.

Christina M. Sanzari, Sasha Gorrell, Lisa M. Anderson, Erin E. Reilly, Martha A. Niemiec, Natalia C. Orloff, Drew A. Anderson, and Julia M. Hormes. The impact of social media use on body image and disordered eating behaviors: Content matters more than duration of exposure. *Eating Behaviors*, 49:101722, April 2023. ISSN 1471-0153. doi: 10.1016/j.eatbeh.2023.101722. URL https://www.sciencedirect.com/science/article/pii/S1471015323000223.

Jose Ramon Saura, Ana Reyes-Menendez, and Stephen B. Thomas. Gaining a deeper understanding of nutrition using social networks and user-generated content. *Internet Interventions*, 20:100312, April 2020. ISSN 2214-7829. doi: 10.1016/j.invent.2020.100312. URL https://www.sciencedirect.com/science/article/pii/S2214782919300363.

Ashley Sharma and Carol Vidal. A scoping literature review of the associations between highly visual social media use and eating disorders and disordered eating: a changing landscape. *Journal of Eating Disorders*, 11(1):170, 2023.

William M Silberg, George D Lundberg, and Robert A Musacchio. Assessing, controlling, and assuring the quality of medical information on the internet: Caveant lector et viewor—let the reader and viewer beware. *Jama*, 277(15):1244–1245, 1997.

Kelly Squires, Alisha Brighton, Lisa Urquhart, Lucy Kocanda, and Susan Heaney. Informing online professional dietetics practice: The development and pilot testing of the Social Media Evaluation Checklist. *Nutrition & Dietetics*, 80 (4):351–361, 2023. ISSN 1747-0080. doi: 10.1111/1747-0080.12794. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1747-0080.12794. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1747-0080.12794.

Stöckl, A. Dynamic Topic Modeling of Video and Audio Contributions. doi: 10.3217/978-3-85125-976-6-18. URL https://diglib.tugraz.at/download.php?id=659e4e1bc1bf2&location=datacite.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text Classification via Large Language Models, October 2023. URL http://arxiv.org/abs/2305.08377. arXiv:2305.08377 [cs].

Shabbir Syed-Abdul, Luis Fernandez-Luque, Wen-Shan Jian, Yu-Chuan Li, Steven Crain, Min-Huei Hsu, Yao-Chin Wang, Dorjsuren Khandregzen, Enkhzaya Chuluunbaatar, Phung Anh Nguyen, and Der-Ming Liou. Misleading Health-Related Information Promoted Through Video-Based Social Media: Anorexia on YouTube. *Journal of Medical Internet Research*, 15(2):e2237, February 2013. doi: 10.2196/jmir.2237. URL https://www.jmir.org/2013/2/e30.

Hao Tang, Sungwoo Kim, Priscila E Laforet, Naa-Solo Tettey, and Corey H Basch. Loss of weight gained during the covid-19 pandemic: Content analysis of youtube videos. *JMIR Formative Research*, 6(2):e35164, 2022.

Jason Thies, Lukas Stappen, Gerhard Hagerer, Björn W. Schuller, and Georg Groh. GraphTMT: Unsupervised Graph-based Topic Modeling from Video Transcripts. In *2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM)*, pages 1–8, November 2021. doi: 10.1109/BigMM52142.2021.00009. URL https://ieeexplore.ieee.org/abstract/document/9643345.

J Kevin Thompson, Leslie J Heinberg, MN Altabe, and S Tantleef-Dunn. Theory assessment, and treatment of body image disturbance. *Thomson JK, Heinberg LJ, Altabe MN, Tantlee-Dunn. Exacting beauty: theory, assessment, and treatment of body image disturbance. Washington, DC: American Psychological Association*, 1999.

Marika Tiggemann and Mia Zaccardo. "Exercise to be fit, not skinny": The effect of fitspiration imagery on women's body image. *Body Image*, 15:61–67, September 2015. ISSN 1740-1445. doi: 10.1016/j.bodyim.2015.06.003. URL https://www.sciencedirect.com/science/article/pii/S1740144515000893.

Marika Tiggemann, Owen Churches, Lewis Mitchell, and Zoe Brown. Tweeting weight loss: A comparison of #thinspiration and #fitspiration communities on Twitter. *Body*

Image, 25:133–138, June 2018. ISSN 1740-1445. doi: 10.1016/j.bodyim.2018.03.002. URL https://www.sciencedirect.com/science/article/pii/S1740144517305375.

Michael A. Traver, Corey M. Passman, Timothy LeRoy, Leah Passmore, and Dean G. Assimos. Is the Internet a Reliable Source for Dietary Recommendations for Stone Formers? *Journal of Endourology*, 23(4):715–717, April 2009. ISSN 0892-7790. doi: 10.1089/end.2008.0490. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2827241/.

Jean E. Fox Tree. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language*, 34(6):709–738, December 1995. ISSN 0749-596X. doi: 10.1006/jmla.1995.1032. URL https://www.sciencedirect.com/science/article/pii/S0749596X85710327.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models, March 2023a. URL http://arxiv.org/abs/2203.11171. arXiv:2203.11171 [cs].

Yuqing Wang, Yun Zhao, and Linda Petzold. Are Large Language Models Ready for Healthcare? A Comparative Study on Clinical Language Understanding. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, pages 804–823. PMLR, December 2023b. URL https://proceedings.mlr.press/v219/wang23c.html. ISSN: 2640-3498.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837, December 2022. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Zihao Wu, Lu Zhang, Chao Cao, Xiaowei Yu, Zhengliang Liu, Lin Zhao, Yiwei Li, Haixing Dai, Chong Ma, Gang Li, Wei Liu, Quanzheng Li, Dinggang Shen, Xiang Li, Dajiang Zhu, and Tianming Liu. Exploring the Trade-Offs: Unified Large Language Models vs Local Fine-Tuned Models for Highly-Specific Radiology NLI Task. *IEEE Transactions on Big Data*, 11(3):1027–1041, June 2025. ISSN 2332-7790. doi: 10.1109/TBDATA.2025.3536928. URL https://ieeexplore.ieee.org/abstract/document/10887002.

Michelle Zeng, Jacqueline Grgurevic, Rayan Diyab, and Rajshri Roy. #WhatIEatinaDay: The Quality, Accuracy, and Engagement of Nutrition Content on TikTok. *Nutrients*, 17(5):781, February 2025. ISSN 2072-6643. doi: 10.3390/nu17050781. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11901546/.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Genan Dai, Nan Yin, Yangyang Li, and Liwen Jing. Investigating Chain-of-thought with ChatGPT for Stance Detection on Social Media, October 2024. URL http://arxiv.org/abs/2304.03087. arXiv:2304.03087 [cs].

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT, March 2023. URL http://arxiv.org/abs/2302.10198. arXiv:2302.10198 [cs].

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models, April 2023. URL http://arxiv.org/abs/2205.10625. arXiv:2205.10625 [cs].