Educational Content on YouTube: The Case of Data Systems

What Engages Learners, and What Educates Them?

MSc. Thesis Computer Science

Xiaojun Ling

Delft University of Technology

Educational Content on YouTube: The Case of Data Systems

What Engages Learners, and What Educates Them?

by

Xiaojun Ling

to obtain the degree of Master of Science
in Computer Science
at the Delft University of Technology,
to be defended publicly on Thursday July 24, 2025, at 3:00 PM.

Student number:	5907756	
Project duration:	December 18, 2024 – July 24, 2025	
Thesis committee:	Dr. ir. MS (Sole) Pera,	TU Delft, Thesis advisor
	Dr. ir. E. Aivaloglou,	TU Delft, Daily Supervisor
	Dr. ir. A. Katsifodimos,	TU Delft, Committee Member
Collaborator:	Dr. ir. M. Liut,	University of Toronto
	Y. Noviello, PhD candidate,	TU Delft

TU Delft

Preface

Embarking on this thesis journey has been both intellectually challenging and personally transformative. From the initial conceptualization to the final stages of writing, this project has taught me to navigate uncertainty, refine focus amidst complexity, and persevere through moments of doubt.

I am deeply grateful to my daily supervisor, Dr. Fenia Aivaloglou, for her continuous, patient guidance throughout the entire process. Her timely support ensured that the work progressed steadily, and her clear, thoughtful feedback consistently helped me regain clarity and focus when I found myself overwhelmed by tangled ideas or anxious emotions. Her mentorship has been invaluable to this thesis.

I am also sincerely grateful to my main advisor, Dr. Sole Pera, whose insights and critical feedback improved the research methods and strengthened the validity of the results. Her insightful questions consistently challenged me to think more deeply and enhanced the rigor of the research. I extend my thanks to Dr. Michael Liut for providing an essential anchor of the research, as well as for his continued input and guidance on the research direction. My appreciation also goes to Yuri Noviello, whose feasible advice on the research approach and technical guidance contributed to the development of many key components of this project.

To my friends, thank you for walking alongside me through the highs and lows of this journey. Your companionship and encouragement, offered even as you faced your own challenges, have been a source of stability and motivation. The mutual support we've shared, especially during moments of difficulty, reminded me of the strength found in empathy and shared resolve.

Lastly, I want to express my heartfelt thanks to my parents. Throughout almost two decades of education, they have been my steady source of support and grounding. No matter where I am or what lies ahead, their presence has always made it possible for me to keep moving forward.

Xiaojun Ling
Delft, July 2025

Abstract

The advancement of data systems demands continuous learning, yet traditional educational materials often fall short of meeting evolving learning needs. YouTube has emerged as a widely used platform for informal learning, but its role in data systems education remains underexamined. This thesis addresses that gap by constructing a curated dataset of 17,434 instructional YouTube videos related to data systems, focusing on content availability and organization, key video characteristics, factors influencing audience engagement, and subtopic coverage in SQL education. Using a curriculum-aligned query strategy and a machine learning filtering pipeline, the dataset maintains relevance to educational objectives and offers broad topical coverage within the data systems domain. Video characteristics are analyzed across dimensions such as content volume, engagement metrics, transcript availability, language, geographic origin, and topic distribution. Findings reveal that content and engagement are highly uneven, with a small subset of videos, channels, languages, countries, and topics capturing disproportionate attention. Statistical modeling shows that engagement in this domain is positively associated with longer video duration, SQL focus, and high-subscriber channels, while overly long titles, frequent uploads, and older channels correlate negatively. Subtle patterns suggest that culturally or regionally tailored content may further enhance engagement. While SQL-related topics dominate in volume and engagement, a subtopic classification of 4,242 SQL videos reveals that although 87% of textbook-derived subtopics are covered, content is heavily concentrated on core querying and schema commands. Advanced, theoretical, systems-level, and integration-related topics are rarely addressed.

Contents

1 Introduction

Data systems education has become increasingly vital in both industry and academia, as organizations in industry rely on robust data management and database skills in the era of big data and Data systems education has long been an essential element of various information technology programs in higher education [1, 90]. However, formal educational offerings in this domain often struggle to keep up with the rapid evolution of technology. Recent studies indicate a misalignment between the competencies demanded by industry and the content taught in academic programs, with curricula adapting slowly to emerging needs [58].

At the same time, online platforms have risen to fill this gap by offering more up-to-date and accessible learning materials. YouTube, in particular, has emerged as an important informal learning platform for technical subjects, including computer science and data systems. According to recent reports, YouTube is one of the most visited websites globally, with over 2.7 billion monthly active users and more than 1 billion hours of video watched daily[1]. It has provided education with a large collection of content spanning diverse academic disciplines, transforming the way people engage with learning materials. Extensive studies have demonstrated YouTube's broad role in education, from informal [74] to formal education [64], and from self-directed learning [52] to integration in classroom teaching [29]. YouTube is also used to enhance traditional education, often with positive outcomes. For instance, engineering students report that certain YouTube channels provide mobile, multi-functional support that traditional classroom settings fail to offer [22]. These examples highlight YouTube's growing influence on learner engagement and comprehension.

Despite YouTube's popularity as a learning tool, its educational content in specialized fields remains fragmented across topics and creators. Many studies of YouTube in education focus on general educational use [79, 63, 54]. Field-dedicated studies remain limited or tend to be narrowly scoped, often concentrating on a single video series, channel, or pedagogical module [8, 22, 70, 93], rather than offering a broader mapping of the landscape. This is especially true in computing domains like data systems, where comprehensive studies that measure the YouTube content in the domain are still lacking. We also know relatively little about what drives audience engagement for educational videos in such specialized fields. Prior research in science communication studied whether factors like how content is structured and presented [89] and the emotional cues [17, 24] in delivery can influence viewer engagement. However, it remains unclear which features make data systems videos engaging, and to what extent the content actually provides coverage of topics in the domain. In summary, there is a need to map out the landscape of data systems educational content on YouTube and to investigate what engages learners versus what truly educates them.

To address these gaps, this thesis is guided by the following research questions:

1. What types of educational video content related to data systems are available on YouTube, and how can they be systematically collected and organized?

[1] https://blog.youtube/press/

2. What are the key characteristics of educational YouTube videos on data systems in terms of content volume, engagement metrics, transcript availability, language, geographic distribution, and topical coverage?
3. Which features of educational YouTube videos on data systems are associated with audience engagement?
4. What gaps and popular areas exist in YouTube video coverage of data systems topics, based on alignment with academic textbook-derived topics?

To answer RQ1, the thesis constructs a curriculum-aligned dataset of educational YouTube videos on data systems using structured search queries derived from international curriculum surveys and a multi-level data collection pipeline. This includes metadata extraction, transcript retrieval, and a relevance filtering process combining manual annotation, large language model classification, and supervised learning. RQ2 is addressed through descriptive analysis of the curated dataset, examining trends in video content volume, engagement metrics, transcript availability, language usage, geographic distribution of channels, and topical coverage. For RQ3, the thesis models audience engagement by identifying and analyzing structural, linguistic, and contextual attributes derived from video and channel-level metadata associated with user interaction, using correlation analysis and explanatory modeling to uncover significant engagement factors. RQ4 is answered by an in-depth case study on Structured Query Language (SQL), a core subfield in data systems education. Using a textbook-derived structure of SQL subtopics, the thesis classifies relevant videos into fine-grained categories with large language models, making it possible to identify both commonly addressed areas and critical gaps in YouTube's educational coverage of SQL.

This work makes several contributions: (1) it presents an open dataset[2] of $17,434$ data systems educational videos on YouTube with comprehensive metadata such as engagement metrics, channel information, comments, and transcripts; (2) it provides empirical insights through detailed descriptive analyses that map the global landscape of data systems content on YouTube, including patterns in volume, engagement, transcript coverage, language use, and geographic distribution; (3) it develops models to identify which video attributes are associated with stronger audience engagement, offering new understanding of what makes educational videos in data systems field more impactful; and (4) it analyzes topic coverage by classifying YouTube videos against a textbook-derived structure, revealing both strengths and gaps in the platform's curriculum coverage and pointing to opportunities for future content development.

The structure of this thesis is organized to address the research questions and objectives in a logical progression. Chapter 2 reviews existing research on YouTube as an educational platform and surveys the current state of data systems education. Chapter 3 describes the dataset creation process, including data collection, filtering, and descriptive analysis, to address the first two research questions concerning content availability and characteristics. Chapter 4 focuses on audience engagement, examining factors associated with engagement (RQ3), explaining the feature engineering and modeling approaches used to analyze engagement, and presenting the results of correlation analyses and modeling. Chapter 5 addresses RQ4 by investigating how well YouTube's educational videos cover the range of textbook-derived topics. Using SQL topics as a case study, it outlines the extraction of subtopics from authoritative database textbooks, the classification of YouTube videos into these subtopics, and the analysis of coverage gaps versus popular content areas. Chapter 6 discusses the findings, their implications for educators and learners, and chapter 7 explains the limitations of the study. Finally, chapter 9 summarizes the contributions of the thesis and proposes directions for future research.

[2] https://doi.org/10.17605/OSF.IO/FTN2S

2
Related Work

This chapter reviews prior research to contextualize the thesis within two key areas: YouTube as an educational platform and the state of data systems education. Section 2.1 surveys interdisciplinary studies that examine YouTube's role in informal and formal education, emphasizing STEM fields and identifying research gaps in domain-specific content. Section 2.2 then reviews the literature on data systems education, highlighting both instructional challenges and pedagogical innovations in foundational topics of data systems.

2.1. YouTube as an Educational Platform

Existing literature reviews have provided broad interdisciplinary insights into the educational applications of YouTube. For instance, Dughera et al. [19] offer a comprehensive synthesis of research on YouTube and its relationship with teaching and learning practices, identifying four major thematic strands: general learning via YouTube, users' educational needs and motivations, the platform's role in formal education, and the practices of educational content creators, or "edutubers". Shoufan and Mohamed [79] similarly synthesized research on YouTube in education, identifying general themes including content creation, user attitudes, usage behaviors, and learning outcomes. Snelson [81] provided an early review in 2011 of YouTube research by that time, summarizing research distribution across different disciplines, pedagogical application methods, and research trends.

Despite a growing body of research on YouTube in educational contexts, most of it remains fragmented within disciplines. In Science, Technology, Engineering, and Mathematics (STEM) education, many studies focus on isolated use cases within particular courses or classroom environments. Comprehensive investigations that characterize educational content on YouTube across an entire academic field, such as computer science or even more specific data systems, still remain scarce. Outside of STEM, Duncan et al. [20] conducted a study in the domain of clinical skills education, evaluating 100 YouTube videos across ten common nursing procedures using structured criteria for both pedagogical and technical quality.

Across several STEM disciplines, particularly Mathematics, Physics, Information Technology, Electrical Engineering, and Mechanical Engineering, there is a shared focus on evaluating YouTube's potential as an educational tool to improve learning outcomes. This interest is pronounced in Mathematics and Engineering, where many studies investigate the platform's role in enhancing student engagement, comprehension, and satisfaction.

For instance, in Mathematics education, studies have shown that YouTube facilitates self-paced learning, increases access to explanations, and fosters learner autonomy [52]. A study found that engineering students view mathematics YouTube channels as mobile and multifunctional sources of help that traditional settings often fail to offer [22]. Similarly, Insorio and Macandog highlighted how teacher-created YouTube video lessons enhanced student understanding and performance in modular distance learning environments [35]. Sari et al. observed that while students initially used YouTube more for entertainment, integrating educational content transformed their experience of mathematics into a more enjoyable and comprehensible process [75]. Additionally, research further emphasized the value of YouTube mathematics lectures in Content and Language Integrated Learning (CLIL) settings, showing their effectiveness in developing both domain knowledge and academic English communication skills among mathematics students [25].

In Engineering disciplines, particularly Electrical and Mechanical Engineering, research has emphasized YouTube's value in conveying abstract and technical concepts. Studies conducted qualitative and quantitative studies on an Electrical Engineering YouTube channel, revealing high perceived educational value and student engagement, especially for complex content that benefits from multimedia explanations [49, 50]. Kibirige and Odora showed that YouTube videos improved cognitive achievement among technology students in a mechanical systems module compared to traditional PowerPoint lectures [40]. Kanetaki et al. demonstrated how YouTube could support sustainable, hybrid learning in Mechanical Engineering, offering consistent engagement across both remote and in-person learning contexts [38].

In Automotive Engineering, studies have focused on developing and evaluating YouTube-assisted learning models tailored to improve students' grasp of complex procedures, such as those involved in engineering drawing. Haryanto et al. [29] proposed a video-assisted project-based learning model using YouTube videos, specifically curated and integrated into a structured guidebook. These videos, including animated tutorials and 2D/3D projection simulations, visually simplify intricate drawing processes and foster independent, flexible learning. The model demonstrated high feasibility and received positive responses from educators and students.

Research in Environmental Science primarily investigates how YouTube videos engage public audiences and shape perceptions of environmental issues by leveraging various strategies such as real-time documentation, influencer collaboration, and event-based educational content [97, 66, 39].

In the broader context of science communication, several studies have explored the factors influencing the communicative effectiveness and user engagement of science-related content on YouTube. This includes the impact of communication styles such as humor and aggression on message reception and activism intentions [101]; the interplay between cognitive features (e.g., segmenting and signaling) and user engagement metrics [89]; the role of emotional cues and affective language in shaping audience responses [24, 17]; and the orchestration of multimodal ensembles, such as embodied, linguistic, and filmic modes, to enhance clarity, coherence, and engagement [93, 9]. Scholars have also examined how informal educational videos reach diverse audiences globally [8], the implications of comment sentiment, participatory behaviors, and argumentative expressions [17, 46], and how references to popular culture like movies may both support and distract from conceptual understanding [46]. Broader analyses of popular science YouTubers and educational content production further highlight success factors such as video structure, editing styles, and educational intent [95, 65].

Complementing these content- and audience-centered studies, other research has examined the perspectives and practices of science communicators themselves. For instance, some studies explore how prominent YouTubers perceive their audience relationships, assess their societal impact, and navigate the algorithmic constraints of the platform [30]. Others focus on the sociodemographic profiles, motivations, financial sustainability, and institutional affiliations of science content creators, revealing an ecosystem that includes both individual and organizational actors with varied goals, resources, and levels of audience engagement [15].

A few works have adopted approaches more closely aligned with our study, either by analyzing large numbers of existing YouTube videos or by applying automated filtering and classification techniques. Kadriu et al. [36] conducted a large-scale investigation of programming-related YouTube tutorials, scraping thousands of videos and analyzing their metadata using a quantitative framework. Their work focused on instructional trends such as programming language popularity, publication timelines, and viewer engagement metrics. They also examined video localization and presentation language, aiming to characterize patterns in self-directed learning through YouTube. Several studies have developed YouTube video classification methods using machine learning techniques. For instance, Kalra et al. [37] applied natural language processing and Random Forest classifiers to categorize approximately 6,000 YouTube videos into six general categories, such as Travel, Food, and Art, based on their titles and descriptions. Meanwhile, Ajwani and Arolkar [5] focused specifically on classifying computer science-related videos. They used keyphrase extraction from VTT (Video Text Track) files and a Random Forest model trained on keyphrase weights to assign videos to specialized areas like Artificial Intelligence and Bioinformatics. Their approach relies on the Computer Science Ontology (CSO), a broad taxonomy well-suited for general CS research domains, but not necessarily aligned with the finer-grained and pedagogically grounded categories needed for data systems and SQL instruction. Generally, while these studies share methodological overlap with our work in using text-derived features to perform large-scale filtering or categorization, their classification targets remain high-level and often coarse-grained. Our study implements a domain-specific filtering mechanism tailored to data systems, followed by a fine-grained subtopic classification aligned with SQL learning objectives, which are not addressed in those works. Additionally, a research [14] similarly approached YouTube at scale, but with a focus on assessing video quality rather than content area. They experimented with multiple machine learning models to predict whether videos are reputable or not, based on attributes such as view counts, comment counts, title length, and sentiment polarity.

2.2. Data Systems Education

Research that examines data systems education as an integrated whole remains limited. Much of the existing literature tends to focus on isolated components such as SQL instruction, database design, or individual pedagogical interventions. Miedema et al. [58] recently presented a structured synthesis that combines curriculum guidelines, course syllabi, and industry input to provide a comprehensive overview of the data systems education landscape. It reveals a misalignment between industry needs and academic offerings.

More research focuses on specific foundational topics within data systems. One such area is conceptual modeling, which is a critical step in database design. Several studies have focused on improving the teaching of conceptual Entity-Relationship (ER) modeling by addressing student difficulties and developing instructional strategies. Research has examined why novices struggle with modeling tasks, emphasizing issues such as misinterpreting assumptions, difficulties in identifying relationships, and cognitive overload during abstraction and semantic transformation [32, 78, 34]. To address these challenges, various methods have been proposed, including the use of concept maps [78], interactive learning systems [34], and structured modeling templates or worksheets to organize problem information [99, 10]. These approaches aim to reduce cognitive load, improve contextual understanding, and support accurate model construction through guided practice and reusable patterns.

Research on database design education has explored how to enhance both conceptual understanding and student engagement through diverse pedagogical approaches. Project-based learning and constructivist models have been used to connect theory to practice and foster authentic, collaborative design experiences [16, 12, 42]. To support novice designers, researchers have developed intelligent or visual tools that guide normalization and schema development [7, 18]. Gamified techniques have also been introduced to increase motivation and reinforce normalization principles through level-based progression [18]. Some studies examine which factors make database assignments engaging for students, identifying personal interest, perceived real-world relevance, and well-matched structural complexity as key contributors [57, 85].

Among all the query languages, SQL is the most widely taught and studied one in data systems education, forming a core component of nearly all database-related curricula. As such, it has become a primary focus of research seeking to improve student learning outcomes, engagement, and instructional efficiency. Researchers have compared the relative difficulty of different SQL constructs [3]. Some have then investigated the types, causes, and persistence of student errors, especially semantic, logical, and syntax errors across various query concepts and assignment formats [4, 86, 100]. Knowledge transfer across query languages is also discussed [48]. Some of the works have examined underlying misconceptions or recognized the importance of it, both from expert interpretations and students' perspectives, to inform more targeted instructional strategies [56, 87]. In response, pedagogical interventions include visual query representations, semantic modeling tools, and conversational agents to support understanding and reduce error rates [55, 45, 68]. Gamification has emerged as another approach for improving motivation and engagement [6, 60]. Moreover, some researchers have proposed the automated generation of SQL exercises using large language models to scale and personalize practice [2]. Broader reviews of SQL education also highlight gaps in teaching methods and recommend research-driven frameworks for improving curricular design and learner support [87, 72].

Understanding how queries are executed in database systems is critical for advanced data systems education. To support this, pedagogical frameworks and instructional modules have been proposed to scaffold the learning of relational query optimization [55]. Various studies have focused on improving learners' comprehension of query execution plans (QEPs). Tools have been developed to visualize and explain QEPs in intuitive ways, allowing students to explore how alternative physical operators affect execution strategies and performance [88, 96]. Some systems translate QEPs into natural language to enhance interpretability and accessibility for novices [51, 98]. Others highlight the potential of QEPs for revealing semantic errors that are not flagged by syntax-based checks, advocating their educational use for usability improvements and deeper understanding of query logic [84].

Database programming education focuses on bridging theory with practical application through various pedagogical methods. Active learning strategies, such as problem-driven labs and "learning by doing," have been implemented to help students gain hands-on competence in SQL and PL/SQL [92]. To improve student engagement, studies have explored fun and interactive approaches, including gamified learning platforms and simulations of real-world vulnerabilities like SQL injection [69, 76]. Some researchers have focused on integrating security practices directly into database programming curricula via modular gamified systems that teach secure coding principles such as input validation and authentication [83]. Others have proposed frameworks that embed database access programming into web development workflows, supported by tools like continuous integration and automated testing to enhance code quality and learning outcomes [67, 53]. These efforts reflect a trend toward practical, security-aware, and student-centered instruction in database programming courses.

3

Data Systems Educational Content on YouTube

This chapter describes how the dataset of YouTube educational videos was constructed and analyzes its main characteristics. Section 3.1 focuses on the methods used to obtain the dataset, detailing the processes of data collection and preprocessing. Section 3.2 then presents the results for RQ1 by describing the dataset's composition and structure, and addresses RQ2 through a descriptive analysis of key characteristics of the dataset, including content volume, engagement metrics, transcript availability, language distribution, geographic origin, and topical coverage.

3.1. Methods

To answer *RQ1: What types of educational video content related to data systems are available on YouTube, and how can they be systematically collected and organized?*, we developed a data collection and processing process as shown in Figure 3.1. This section details our methods of dataset construction, including search query design, multi-level metadata extraction, transcript retrieval, preprocessing, and a relevance filtering process. The process was designed to ensure both breadth of topical coverage and precision in isolating genuinely instructional content, enabling subsequent analyses of educational trends and content characteristics on the platform.

3.1.1. Dataset Construction

Topic Query Design

To construct a dataset of educational YouTube videos relevant to data systems, we needed a comprehensive and representative set of search queries. We grounded our topic selection in the survey presented by Miedema et al. [58]. The research synthesizes input from 19 national and international curriculum guidelines across 12 countries and global organizations, integrating both academic and industry perspectives. The study surveyed 105 post-secondary data systems educators from 24 countries, as well as 34 industry professionals, to identify the knowledge areas and competencies most relevant to both formal education and professional practice. In the educator survey in the study, respondents were asked to report topic coverage in their courses based on a structured list of high-level topics and 38 detailed subtopics. These subtopics presented in Table 2 of the report served as the foundation for defining our search scope. For each subtopic, we created one or more natural language search queries designed to retrieve relevant content using the YouTube Data API. Table 3.1 provides a mapping between the original subtopic formulations and the corresponding search queries used in our dataset collection process. In total, we covered all 38 subtopics from the educator survey in the reference study, converting them into 67 distinct search queries. This approach balances coverage and precision, aiming to retrieve a diverse but thematically relevant set of videos across both general and specific instructional content relevant to each concept.

By anchoring the topic queries in this structured and internationally informed curriculum framework, we

Survey Subtopic(s)	**Expanded Search Queries**
relational theory: relations, tuples and attributes	”relational theory”, ”relational theory relations”, ”relational theory tuples”, ”relational theory attributes”
tuple relational calculus	”tuple relational calculus”
relational algebra	”relational algebra”
data visualization	”data visualization”
database optimization: indexing	”database optimization”, ”database optimization indexing”
database optimization: query execution plans	”database optimization query execution plans”
database optimization: query optimization	”database optimization query optimization”
database scalability: replication	”database scalability”, ”database scalability replication”
database scalability: sharding	”database scalability sharding”
NoSQL database management systems	”NoSQL database management systems”
logical and physical data independence	”data independence”, ”logical data independence”, ”physical data independence”, ”logical and physical data independence”
database management system components	”database management system components”
functions and stored procedures	”functions and stored procedures”
data modeling: conceptual modeling	”data modeling”, ”data modeling conceptual modeling”
data modeling: mapping conceptual models to logical models	”data modeling mapping conceptual models to logical models”
data modeling: creating tables and columns	”data modeling creating tables and columns”
database normalization: functional dependency	”database normalization”, ”database normalization functional dependency”
database normalization: candidate	”database normalization candidate”
database normalization: super keys	”database normalization super keys”
database normalization: normal forms up to BCNF	”database normalization normal forms up to BCNF”
database normalization: multivalued dependency	”database normalization multivalued dependency”
database normalization: join dependency	”database normalization join dependency”
object-oriented data models	”object-oriented data models”
semi-structured traditional data models	”semi-structured traditional data models”
SQL: select, project, join	”SQL”, ”SQL select”, ”SQL project”, ”SQL join”
SQL: insert, update, delete	”SQL insert”, ”SQL update”, ”SQL delete”
SQL: aggregation and group by	”SQL aggregation”, ”SQL group by”
SQL subqueries	”SQL subqueries”
SQL: common table expressions	”SQL common table expressions”
transaction processing	”transaction processing”
concurrency control and isolation levels	”concurrency control”, ”isolation levels”, ”concurrency control and isolation levels”
database back-ups and recovery	”database back-ups”, ”database recovery”, ”database back-ups and recovery”
distributed database management systems	”distributed database management systems”
data mining: algorithm	”data mining”, ”data mining algorithms”
data mining: associative and sequential patterns	”data mining associative pattern”, ”data mining sequential pattern”, ”data mining associative and sequential patterns”
data mining: data cleaning	”data mining data cleaning”
data mining: market basket analysis	”data mining market basket analysis”
data privacy and ethics	”data privacy”, ”data ethics”, ”data privacy and ethics”
data security and database access management	”data security”, ”database access management”, ”data security and database access management”
data warehousing	”data warehousing”

Table 3.1: Mapping from survey subtopics [58] to YouTube API search queries

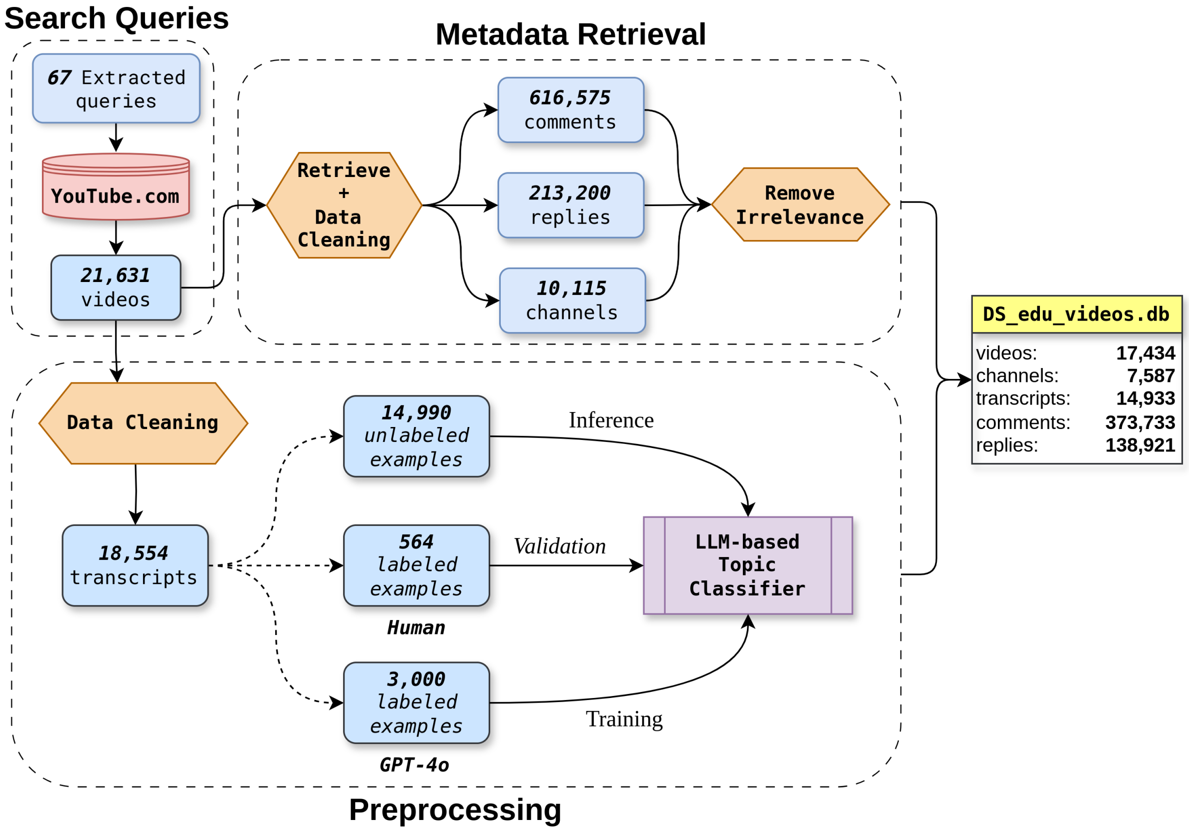


Figure 3.1: Overview of the dataset construction process

ensure that the resulting dataset reflects the actual instructional needs and terminologies of the global data systems education community. This also facilitates downstream analyses of content coverage and topic representation.

Metadata Collection

Video metadata was collected using the YouTube Data API v3[1] through a staged querying process. For each search query derived from the curriculum-informed expert curated topic list, videos were retrieved in descending order of view count using paginated calls to the `search.list` endpoint, with up to 50 results per page. From each response, we extracted basic metadata including the video ID, title, channel name, and publication date.

Subsequently, we passed the collected video IDs to the `videos.list` endpoint to retrieve detailed video-level metadata. This included:

- Content attributes: video description, tags, duration, and definition (HD/SD);
- Engagement statistics: view count, like count, comment count;
- Language and accessibility: default audio language, default textual language, caption availability;
- Monetization flag: the presence of paid product placement.

The initial video search and metadata collection were completed on December 24, 2024, yielding a dataset of $21,631$ videos. In the following week, a secondary round of data collection was conducted to retrieve channel-level and comment-level metadata for these videos.

To enrich video-level data with additional context, we extracted channel-level metadata by resolving channel IDs through the `channels.list` endpoint. This included subscriber count, total view count, upload count, country of origin, and channel creation date, along with both default and English-localized

[1] `https://developers.google.com/youtube/v3`

channel titles and descriptions. Each video entry was linked back to its source channel for potential cross-level analysis. In total, we successfully retrieved channel metadata for $10,115$ entries.

Finally, we collected comment-level metadata for each video using the `commentThreads.list` and `comments.list` endpoints, handling pagination to ensure all comments are retrieved. To facilitate relational analysis, top-level comments and their replies were stored in two separate normalized tables, `comments` and `replies`, respectively, within the same database.

Each top-level comment was stored in the `comments` table, which included the comment text, number of likes, publication and last update timestamps, and the total number of replies. These entries were uniquely identified by a thread ID and linked to their parent video via a foreign key. All replies to these top-level comments were stored in the `replies` table, with each reply associated with both its thread and video. The table similarly captured the reply text, like count, and timestamps. This structure supports many-to-one relationships from replies to comments and enables precise mapping of threaded discussions under each video. In total, we collected $616,575$ top-level comments and $213,200$ replies.

During the secondary round of data collection, some videos became unavailable due to deletion or changes in access permissions. After removing these entries, the dataset is reduced to $21,566$ videos.

State-saving mechanisms implemented in scrapers enabled the seamless resumption of scraping after interruptions or quota limits in the process of data collection, ensuring dataset completeness. All metadata was stored in a structured SQLite database.

Transcript Collection

To support downstream text-based analyses, we collected English transcripts for each video using the `youtube-transcript-api` Python library[2]. The transcript retrieval was performed separately from the metadata collection, using the list of video identifiers previously stored in the dataset. All transcripts were fetched via an authenticated proxy service to improve request reliability and access rate. The process of transcript collection was performed in two phases.

In the initial phase, we attempted to retrieve English transcripts directly for each video ID in the dataset. The `YouTubeTranscriptApi.list(video_id)` method was used to provide available caption tracks for each video, and the transcript marked for the ”en” language was selected where present. YouTube provides two types of captions: those manually uploaded by the video creator and those automatically generated using speech recognition. Manually uploaded captions are prioritized over auto-generated ones if available. Captions were concatenated into a single string and stored in the SQLite database in a separate transcripts table along with the type and translatability flag of transcripts.

Videos with captions disabled, unavailable, or lacking any transcript were skipped, and retry logic was implemented for transient failures. A persistent state file ensured the process could resume in the event of interruption.

To improve transcript coverage, we conducted an additional pass to identify videos that were missing transcripts in the initial collection phase. For these cases, we attempted to retrieve translated English transcripts, leveraging the built-in translation functionality `transcript.translate('en')`. When a transcript was found in a non-English language, it was translated into English using YouTube’s built-in caption translation service.

In this second phase, we prioritized manually uploaded captions when available, as before. If none were present, we then fell back to automatically generated captions. Translated transcripts were labeled with their type and original language (e.g., ”creator-uploaded (Translated from French)”). This enhanced coverage enabled the inclusion of additional videos in language-based analysis while preserving information about transcript provenance and reliability.

We collected $18,554$ transcripts in total at this stage. The transcripts were stored in a dedicated table `transcripts`, keyed by `video_id`, and linked back to the video metadata tables.

[2] `https://pypi.org/project/youtube-transcript-api/`

3.1.2. Data Processing

Data Cleaning and Preprocessing

Following initial data collection, a series of preprocessing steps were applied to standardize the dataset. While a wide range of features was gathered to ensure analytical flexibility, certain fields were later removed due to redundancy, low variance, or lack of meaningful informational value.

Specifically, the following fields were excluded from the final working dataset:

- `translatable`: The flag indicating whether a transcript could be translated was dropped because all usable transcripts in the final dataset were translatable by default, and it was unnecessary once English-translated transcripts were incorporated.
- `top_level_published_at` and `reply_published_at`: These timestamps were removed because only the most recent update time for each comment and reply to comment was retained, as the original publication time was no longer meaningful without access to full historical text content.
- `likes_playlist`: This field, representing the ID of a playlist containing a channel's liked videos, was consistently empty across all entries and thus discarded.

In addition, all time-related attributes (e.g., video and channel publication dates) were retained in ISO 8601 extended format (e.g., ”2018-06-29T04:52:00Z”) to ensure consistent temporal representation. Video durations, originally encoded in ISO 8601 format (e.g., ”PT6M30S”), were parsed into standard HH:MM:SS strings (e.g., ”0:04:24”) to facilitate human-readable inspection and summary statistics.

These preprocessing steps ensured uniformity across temporal fields and helped reduce feature noise before relevance filtering and feature engineering. All cleaning operations were applied consistently across the video, channel, and comment-level tables.

Relevance Filtering: Sampling and Training Label Production

While all videos in the dataset were retrieved using predefined search queries, YouTube's search engine often returns content that is tangential, promotional, or unrelated to educational goals. For example, the query ”data privacy and ethics” returned videos such as news reports on corporate data breaches and political commentary on surveillance laws, content that, while loosely related, does not serve instructional purposes. Similarly, a query on ”relational theory” produced results about Relational Frame Theory, a concept in behavioral psychology, which is unrelated to relational databases. To ensure the dataset reflected genuinely instructional content in data systems, we implemented a multi-step relevance filtering process combining manual annotation, synthetic training label generation, instruction fine-tuned embedding model, and classifier-based prediction.

We first drew a stratified sample of 564 videos, proportionally distributed across all survey subtopics. Each video was manually annotated as either relevant (i.e., educational and within the scope of data systems) or irrelevant (e.g., marketing, entertainment, or only loosely related). This resulted in 485 relevant and 79 irrelevant samples. These manual labels were used to evaluate automated methods of generating labels for training data.

To expand the labeled set efficiently, we used GPT-4o with designed prompts to classify a broader sample of $3,000$ videos. We tested multiple prompting strategies with GPT-4o, using title, description, and transcript as input, varying two elements: (1) whether to include chain-of-thought (CoT) reasoning and (2) whether to include explicit exclusion criteria concluded based on observations (e.g., excluding news, promotional content, legal interpretations). Each variant was evaluated on the manually labeled sample of 564 videos.

Given the strong class imbalance, where relevant videos (positive class) greatly outnumber irrelevant ones, our priority was to minimize false positives, i.e., avoid retaining off-topic content. We therefore evaluated prompts based on false positive rate (FPR) in addition to accuracy, precision, and recall.

As shown in Table 3.2, the selected prompt without CoT but with exclusion rules achieved a balanced performance with the relatively low FPR (22.8%) among variants that maintained high recall (94.0%) and precision (96.2%). While the With CoT and Exclusion prompt achieved a slightly lower FPR (16.5%), but came with two trade-offs: (1) a noticeable drop in recall (91.6%), meaning more relevant videos were mistakenly excluded; and (2) significantly higher token usage, due to the verbosity introduced by chain-of-thought reasoning. This increases both inference time and computational cost when labeling thousands of samples. The No CoT, No Exclusion variant performed worst on FPR (27.9%) despite slightly higher recall.

Prompt Variant	Accuracy	Precision	Recall	FPR
With Exclusion, No CoT (Selected)	$\mathbf{91.7\%}$	96.2%	94.0%	22.8%
With CoT and Exclusion	90.4%	$\mathbf{97.2\%}$	91.6%	$\mathbf{16.5\%}$
No CoT, No Exclusion	91.5%	95.4%	$\mathbf{94.6\%}$	27.9%

Table 3.2: Performance Comparison of Prompt Variants for Relevance Classification

Given that false positives (i.e., incorrectly retaining irrelevant content) were considered more costly than false negatives in our context, and that cost-efficiency was a practical concern for scaling, the selected prompt offered the best trade-off between precision, recall, and labeling efficiency.

The final chosen prompt included a clear definition of "instructional video", a list of data systems subtopics from the referenced survey, and a set of exclusion criteria. Additionally, the prompt was structured to elicit a strictly binary response, either "1" (relevant) or "0" (irrelevant), to facilitate consistent parsing and automatic label extraction. The full prompt is provided in Appendix section A.1.

Based on the selected prompt, we classified an additional set of $3,000$ videos, which are also sampled proportionally to the video population on different subtopics to ensure topic-wise representativeness and prevent overfitting to more frequent content areas. This expanded set yielded $2,486$ relevant and 514 irrelevant cases, which formed the basis for the synthetic dataset used in downstream classifier development.

Eventually, the labeled data were drawn from two sources: the $3,000$ videos labeled by GPT-4o with the selected prompt (used for training), and the manually annotated set of 564 videos (reserved for validation and testing).

To construct the final training dataset, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to the GPT-labeled subset to balance class distributions. This resulted in a training set with $4,972$ samples, equally split between relevant ($n = 2,486$) and irrelevant ($n = 2,486$) classes. The manually labeled data were stratified and split 50/50 into a validation set and a test set, each containing 282 samples with similar class proportions (approximately 86% relevant and 14% irrelevant). This setup ensured reliable performance evaluation while preventing label leakage between the training and evaluation phases. Table 3.3 summarizes the allocation and class composition. This hybrid dataset provided sufficient coverage of both classes to support supervised filtering while preserving human oversight and LLM-informed generalization.

Data Source	Subset	Relevant (n)	Irrelevant (n)
GPT-4o labeled (3,000 videos)	Training (after SMOTE)	$2,486$	$2,486$
Manual annotation (564 videos)	Validation	243	39
	Test	242	40

Table 3.3: Composition and allocation of labeled data used in classification

Relevance Filtering: Embedding Model Comparison

To scale relevance prediction to the full dataset, we adopted an embedding-based classification method. For each video, we constructed a textual representation using a structured textual prompt composed of three parts: the video title, a keyword-extracted description, and a keyword-extracted transcript. The description and transcript texts were first cleaned and truncated to fit within the model's token limit ($2,000$ tokens for description, $6,000$ for transcript). We applied the RAKE (Rapid Automatic Keyword Extraction) algorithm [73] through the Python package `rake-nltk`[3] to both fields to reduce redundancy and retain key topical phrases. The final composite input took the form:

[3] `https://pypi.org/project/rake-nltk/`

```
Title: <cleaned title>
Description Keywords: <RAKE keywords>
Transcript Keywords: <RAKE keywords>
```

These texts served as input for generating vector embeddings via instruction-tuned language models. This format encouraged the embedding model to focus on semantically rich tokens while discarding noisy or generic language.

To identify suitable embedding models, we initially referred to the Massive Text Embedding Benchmark (MTEB) [62], which ranks models based on their zero-shot and supervised performance across a wide range of NLP tasks. We began with `KaLM-embedding-multilingual-mini-instruct-v1.5` [31], a compact instruction-finetuned model based on `Qwen2-0.5B`, trained using weak supervision followed by supervised fine-tuning. This model was used as an exploratory baseline due to its small size and fast inference, allowing us to quickly validate the feasibility of instruction-tuned embedding models for the task.

Encouraged by the initial performance, we expanded our experiments to two larger and more competitive MTEB-ranked models: `gte-Qwen2-1.5B-instruct` and `gte-Qwen2-7B-instruct` [47]. Each combined input was encoded with an instruction-style prompt `"Instruct: Determine whether the video provides educational content related to data systems."` to align with the instruction-tuned nature of the selected models. Embeddings were generated in batches using the models and were then L2-normalized, stored as NumPy arrays, and indexed by video ID for downstream use in classifier training and full-dataset filtering.

To compare the effectiveness of different embedding models, we trained XGBoost classifiers on the same downstream classification task using embeddings generated by each candidate model. Model performance of the test set was evaluated using a set of metrics:

- ROC-AUC (Receiver Operating Characteristic Area Under the Curve), which measures the model's ability to distinguish between relevant and irrelevant videos regardless of threshold;
- PR-AUC (Precision–Recall AUC), which is more sensitive to class imbalance and highlights performance in identifying positive (relevant) cases;
- F1-score, Precision, and Recall, reported separately for both the relevant (Rel) and irrelevant (Irr) classes to assess class-specific behavior.

This metric suite was chosen to balance global discriminative ability (ROC-AUC), robustness under class imbalance (PR-AUC), and sensitivity to minority class recall, which is important in filtering out off-topic videos.

As shown in Table 3.4, `gte-Qwen2-7B-instruct` consistently outperformed the smaller models across nearly all metrics. It achieved the highest ROC-AUC (0.937) and PR-AUC (0.985), and showed the strongest performance for the irrelevant class with an F1-score of 0.722, reflecting improved ability to correctly detect and exclude irrelevant content. While `Qwen2-1.5B-instruct` produced the highest recall for the relevant class (0.950), its precision and F1 for the irrelevant class were notably lower. Based on these results, `gte-Qwen2-7B-instruct` was adopted as the final embedding model for subsequent filtering and analysis.

Model	ROC-AUC	PR-AUC	F1-Rel	F1-Irr	P-Rel	P-Irr	R-Rel	R-Irr
KaLM-mini	0.873	0.968	0.872	0.562	0.979	0.409	0.785	0.900
Qwen2-1.5B	0.905	0.977	0.862	0.563	**0.989**	0.400	0.764	**0.950**
Qwen2-7B	**0.937**	**0.985**	**0.942**	**0.722**	0.978	**0.614**	**0.930**	0.795

Table 3.4: Embedding Model Performance on Classification using XGBoost Classifier. Model names are abbreviated. The best-performing metrics are highlighted in bold.

In addition to the main experiments, we conducted a controlled ablation to assess the contribution of the instruction component in prompt-based embedding generation. Specifically, we tested both `gte-Qwen2-1.5B-instruct` and `gte-Qwen2-7B-instruct` on the downstream classification task with and without the instruction specified. As shown in Table 3.5, the effect of including this instruction varied by model. For gte-Qwen2-7B-instruct, including the instruction improved performance for the minority (irrelevant) class, increasing F1 from 0.607 to 0.722. In contrast, the 1.5B model performed slightly better without the instruction.

Model	ROC-AUC	PR-AUC	F1-Rel	F1-Irr	P-Rel	P-Irr	R-Rel	R-Irr
Qwen2-1.5B (w/ instr.)	0.905	0.977	0.862	0.563	0.989	0.400	0.764	0.950
Qwen2-1.5B (no instr.)	0.907	0.985	0.895	0.598	0.976	0.455	0.826	0.875
Qwen2-7B (w/ instr.)	$\mathbf{0.937}$	0.985	$\mathbf{0.942}$	$\mathbf{0.722}$	0.978	$\mathbf{0.614}$	$\mathbf{0.930}$	0.795
Qwen2-7B (no instr.)	0.927	$\mathbf{0.988}$	0.891	0.607	$\mathbf{0.985}$	0.451	0.814	$\mathbf{0.925}$

Table 3.5: Impact of Instruction Prompts on Embedding-Based Classification (XGBoost Classifier)

However, this observation is based on the fact that the results are conditioned on a single phrasing of the instruction. Different wordings, task formulations, or input structuring strategies could yield different outcomes. A more comprehensive analysis of instruction design is beyond the scope of this work, but our findings suggest that prompt style should be treated as a tunable component in embedding-based pipelines.

Relevance Filtering: Classifier Performance Comparison and Filtering Outcome

To identify the most suitable classification model for final filtering, we evaluated a suite of commonly used classifiers, including Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM) with both linear and RBF kernels, Multilayer Perceptron (MLP), and XGBoost. All models were trained using embeddings from gte-Qwen2-7B-instruct and evaluated on the same validation and test splits. Class imbalance was addressed through class weighting or scale_pos_weight, and probability calibration was applied for models not natively supporting probabilistic outputs.

We additionally introduced interaction features and applied standard scaling across all models. To account for the asymmetric cost of classification errors, where false positives (retaining irrelevant content) are considered more detrimental, we adopted a cost-aware decision threshold optimization strategy. Specifically, for each model, we selected the classification threshold that minimized the total cost:

$$\text{Total Cost} = \text{FN} \cdot \text{Cost}_{\text{FN}} + \text{FP} \cdot \text{Cost}_{\text{FP}} \tag{3.1}$$

where the cost ratio was set to $\text{Cost}_{\text{FP}} = 2$, $\text{Cost}_{\text{FN}} = 1$.

Table 3.6 reports the model performance on the test set, using standard metrics including ROC-AUC, PR-AUC, F1-scores, precision, and recall for both relevant and irrelevant videos. Among all evaluated models, XGBoost achieved the strongest balance between discrimination power (ROC-AUC = 0.937), recall for relevant videos ($R = 0.930$), and F1 for the minority irrelevant class ($F1 = 0.722$). While MLP achieved slightly higher recall for the relevant class ($R = 0.988$), its F1-score for the irrelevant class was lower (0.590), indicating lower robustness for filtering noise. We therefore selected XGBoost as the final classifier for filtering out irrelevant videos from the dataset.

Table 3.6: Classifier performance using gte-Qwen2-7B-instruct embeddings

Classifier	ROC-AUC	PR-AUC	F1-Rel	F1-Irr	P-Rel	P-Irr	R-Rel	R-Irr
LR	0.924	0.985	0.890	0.588	0.975	0.443	0.818	0.875
RF	0.905	0.976	0.876	0.585	$\mathbf{0.990}$	0.422	0.785	$\mathbf{0.950}$
SVM (Linear)	0.921	$\mathbf{0.986}$	0.914	0.600	0.955	0.500	0.876	0.750
SVM (RBF)	0.921	$\mathbf{0.986}$	0.914	0.600	0.903	0.600	$\mathbf{0.959}$	0.375
MLP	0.926	0.982	$\mathbf{0.950}$	0.590	0.916	$\mathbf{0.857}$	$\mathbf{0.988}$	0.450
XGBoost	$\mathbf{0.937}$	0.985	0.942	$\mathbf{0.722}$	0.978	0.614	0.930	0.795
GPT-4o	-	-	0.950	0.727	0.966	0.667	0.934	0.800

After finalizing the classifier, we applied the trained XGBoost model to the full unlabeled dataset to filter out irrelevant videos. Out of the initial $21,566$ collected videos, $4,132$ were classified as irrelevant and excluded from further analysis. This left a curated dataset of 17,434 videos deemed to be topically relevant and educational in nature. The data entries from other levels referencing the removed irrelevant videos are also removed. The resulting database with 17,434 videos, 7,587 channels, 373,733 comments, 138,921 replies, and 14,933 transcripts forms the final dataset for analyses in subsequent chapters.

3.2. Results

This section presents the results of our data collection and processing process and the descriptive analyses conducted to address both RQ1 and RQ2. Subsection 3.2.1 examines the composition of the curated dataset, details the structure of the resulting relational database, and provides an overview of its key components. Subsection 3.2.2 presents descriptive analysis results organized into five analytical themes: (1) content volume and engagement metrics; (2) transcript availability; (3) language distribution based on audio and textual metadata; (4) geographical distribution; and (5) topic distribution.

3.2.1. RQ1: Dataset Composition

To address *RQ1: What educational content related to data systems is available on YouTube, and how can it be systematically collected and organized?*, we constructed a dataset for data systems educational videos on YouTube. Using the survey from Miedema et al. [58] as a foundational taxonomy, we created 67 keyword queries that expanded 38 high-level data systems subtopics into natural language search strings. These queries were submitted to the YouTube Data API, resulting in the initial retrieval of 21,631 video entries across topics. Metadata was then collected at multiple levels, such as video-level, channel-level, comment-level, and transcript-level. Following metadata and transcript collection, a multi-step filtering pipeline involving embedding models and supervised classification was conducted to exclude off-topic or non-instructional content, resulting in a curated set of 17,434 instructional videos.

To support scalable querying and analysis, all collected data were organized into a normalized relational schema and stored in a SQLite database. The final schema is illustrated in Figure 3.2 as an Entity Relationship (ER) diagram. The database is composed of five interconnected tables:

- `videos`: video-level descriptors and statistics;
- `channels`: channel-level descriptors and statistics;
- `comments`: comment thread with top-level comment resources;
- `replies`: attributes of replies to top-level comments;
- `transcripts`: manually or automatically generated captions associated with each video.

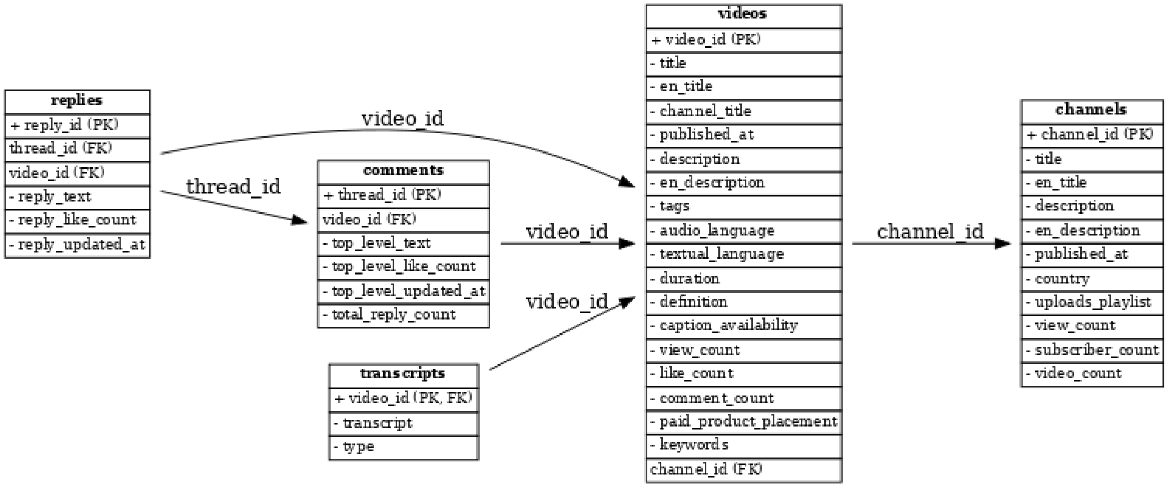


Figure 3.2: Entity Relationship (ER) Diagram of the final YouTube educational video database

Table: `videos` **(17,434)**		
Column	**Description**	**Type**
video_id	Unique YouTube video identifier	TEXT
title	Original video title	TEXT
en_title	English video title (if available, otherwise original title displayed)	TEXT
channel_title	Original channel name	TEXT
published_at	Video publication timestamp (ISO 8601)	TEXT
description	Original video description	TEXT
en_description	English description (if available)	TEXT
tags	List of video tags created by the video uploader to describe the video content	TEXT
audio_language	Language code (I18nLanguages code) of default audio track	TEXT
textual_language	Text content (title, description) language code (I18nLanguages code)	TEXT
duration	Duration in HH:MM:SS format	TEXT
definition	Video resolution (HD/SD)	TEXT
caption_availability	Availability of manual captions (true/false)	TEXT
view_count	Number of views	INTEGER
like_count	Number of likes	INTEGER
comment_count	Number of comments	INTEGER
paid_product_placement	Paid placement indicator (true/false)	TEXT
keywords	List of matched search queries	TEXT
channel_id	Unique YouTube channel identifier	TEXT
Table: `channels` **(7,587)**		
channel_id	Unique channel identifier	TEXT
title	Original channel title	TEXT
en_title	English title (if available)	TEXT
description	Channel description	TEXT
en_description	English description (if available)	TEXT
published_at	Channel creation timestamp (ISO 8601)	TEXT
country	Channel registration country (ISO)	TEXT
uploads_playlist	ID of uploads playlist	TEXT
view_count	Channel total views	INTEGER
subscriber_count	Number of subscribers	INTEGER
video_count	Number of uploaded videos	INTEGER
Table: `comments` **(373,733)**		
thread_id	Unique comment thread identifier	TEXT
video_id	Unique YouTube video identifier	TEXT
top_level_text	Top-level comment text	TEXT
top_level_like_count	Likes on top-level comment	INTEGER
top_level_updated_at	Last update timestamp (ISO 8601)	TEXT
total_reply_count	Number of replies	INTEGER
Table: `replies` **(138,921)**		
reply_id	Unique reply identifier	TEXT
thread_id	Unique comment thread identifier	TEXT
video_id	Unique YouTube video identifier	TEXT
reply_text	Reply text	TEXT
reply_like_count	Likes on reply	INTEGER
reply_updated_at	Last update timestamp (ISO 8601)	TEXT
Table: `transcripts` **(14,933)**		
video_id	Unique YouTube video identifier	TEXT
transcript	Transcript text	TEXT
type	Transcript type (manual / auto-generated)	TEXT

Table 3.7: Overview of the database composition

Foreign key constraints were used to maintain referential integrity between video entries and their associated channels, comments, replies, and transcripts. These relationships support structured joins across metadata levels and enable multi-level analytical operations. The referencing relations are as follows:

- `videos.channel_id` → `channels.channel_id`
- `comments.video_id` → `videos.video_id`
- `replies.thread_id` → `comments.thread_id`
- `replies.video_id` → `videos.video_id`
- `transcripts.video_id` → `videos.video_id`

This schema enables fine-grained analysis at different levels (e.g., per-video engagement, per-channel distribution, or thread-level discourse) and supports efficient downstream processing tasks such as filtering, feature modeling, and topic classification.

Table 3.7 provides an overview of the final dataset structure, including the main tables, their field descriptions, data types, and observed record counts after the full data processing pipeline. The dataset contains $17,434$ curated educational YouTube videos on data systems topics across $7,587$ channels, with $373,733$ top-level comments, $138,921$ threaded replies, and $14,933$ English transcripts. Importantly, this corpus represents the full set of retrievable results from the YouTube Data API for each of the 67 curriculum-aligned search queries, ensuring maximum content coverage within the queryable constraints of the platform. To support further research and reproducibility, the dataset is publicly available at The Open Science Framework[4].

3.2.2. RQ2: Key Characteristics of the Dataset

To address the second research question: *What are the key characteristics of educational YouTube videos on data systems in terms of content volume, engagement metrics, transcript availability, language, geographic distribution, and topical coverage?*, we conducted a set of descriptive analyses based on the final filtered dataset.

We structure this exploration into five analytical facets: (1) content volume and engagement metrics, (2) transcript availability, (3) language characteristics, (4) geographic distribution, and (5) topic coverage trends, reported in the following subsections.

Content Volume and Engagement Metrics

Table 3.8 summarizes descriptive statistics for key video, comment, and channel-level attributes, including quartiles and log-scale histograms, which illustrate the skewed distributions characteristic of online educational content.

Across the dataset, videos span a wide range of durations and audience engagement levels. With a median duration of 10.03 minutes and an interquartile range from 4.82 to 21.96 minutes. However, the substantial standard deviation (47.08 minutes) and maximum duration ($1,753.78$ minutes, approximately 29 hours) indicate a small subset of content creators offering long videos alongside the more prevalent shorter tutorials. The histogram confirms this right-skewed distribution, with most videos clustering in the shorter duration range.

Engagement metrics exhibit classic Power law distribution characteristics. Video view counts show a median of $2,274$ views per video, but with a mean of $45,322.06$ and a maximum of $19,064,477$, indicating that a small percentage of videos capture disproportionate audience attention. This pattern repeats across like counts (median: 35, maximum: $356,316$) and comment counts (median: 2, maximum: $11,494$), which are more sparsely distributed. The near-zero first quartile values for likes (Q1: 6) and comments (Q1: 0) reveal that many data systems educational videos receive minimal engagement, while a select few generate substantial viewer interaction. Comment-level metrics provide additional insights into viewer interaction patterns. The median values of zero for both top-level comment likes and replies indicate that most comments receive no engagement, with interaction concentrated on a small

[4] https://doi.org/10.17605/OSF.IO/FTN2S

subset of comments (maximum likes: $7,839$, maximum replies: 353). Similarly, most replies receive minimal likes (median: 0, mean: 0.83).

The $7,587$ channels in our dataset range from small educational producers (Q1 subscriber count: 158) to major content platforms with up to 42.2 million subscribers. The median channel hosts 121 videos and has accumulated $127,736$ views across its content library. The substantial gap between median and mean values for channel views (median: $127,736$, mean: $8,914,953$) and subscribers (median: $1,280$, mean: $65,692.28$) reinforces the highly skewed nature of the ecosystem, where a few dominant channels have significantly larger audiences than common content creators.

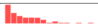
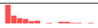
Attribute	Count	STD	Mean	Min	Q1	Median	Q3	Max	Histogram (log scale)
Video Duration (minutes)	$17,434$	47.08	21.37	0	4.82	10.03	21.96	$1,753.78$	
Video View Count	$17,433$	$288,716.23$	$45,322.06$	0	283	$2,274$	$15,081$	$19,064,477$	
Video Like Count	$17,029$	$5,512.79$	859.99	0	6	35	222	$356,316$	
Video Comment Count	$16,948$	187.41	31.40	0	0	2	13	$11,494$	
Comment Like Count	$373,733$	49.43	2.88	0	0	0	1	$7,839$	
Comment Reply Count	$373,733$	2.09	0.43	0	0	0	1	353	
Reply Like Count	$138,921$	6.96	0.83	0	0	0	1	905	
Channel View Count	$7,587$	$126,522,808.75$	$8,914,953$	2	$16,084$	$127,736$	$1,006,418$	$8,187,923,440$	
Channel Subscriber Count	$7,587$	$738,761.20$	$65,692.28$	0	158	$1,280$	$8,600$	$42,200,000$	
Channel Video Count	$7,587$	$29,937.28$	$1,118.51$	1	34	121	351	$2,035,401$	

Table 3.8: Descriptive statistics of key video and channel metrics

Figure 3.3 visualizes the annual trends in video publication and channel creation. Video production rose steadily throughout the 2010s and saw a sharp spike in 2020, coinciding with global shifts toward remote learning due to the COVID-19 pandemic. Since 2021, the number of new videos has plateaued at a relatively high level, indicating sustained content creation post-pandemic.

In contrast, channel creation does not strictly follow the same trajectory. While the number of new educational channels gradually increased in the early years and peaked sharply in 2020, it exhibited notable fluctuations. In particular, there was a smaller surge in 2011. Following this, growth remained relatively steady until the dramatic spike in 2020. Since 2021, the number of new channels has declined significantly, even as video production remained high. This decoupling suggests that recent content growth is driven less by new entrants and more by intensified output from existing creators.

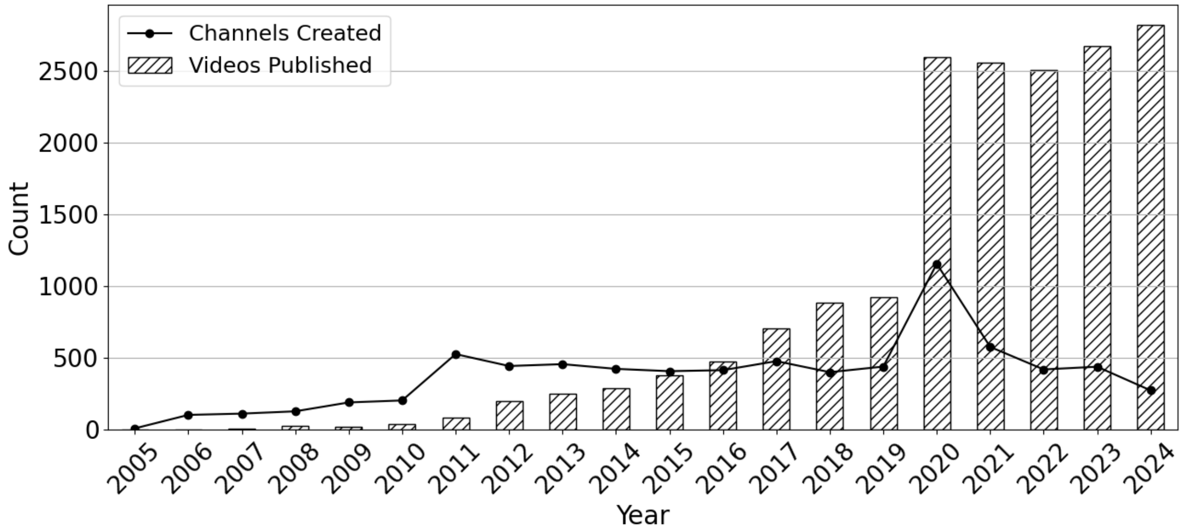


Figure 3.3: Temporal trends of video publishing and channel creation

Transcript Availability

Transcripts are critical for enabling text-based analysis and supporting accessibility and multilingual comprehension. Out of the final set of $17,434$ videos retained after relevance filtering, English transcripts were successfully retrieved for $14,933$ videos, representing an overall coverage of 85.6%. This includes both manually uploaded captions and auto-generated ones provided by YouTube's speech recognition system.

Table 3.9 breaks down the availability and engagement profiles of four transcript types: manually created captions, translations of manual captions, auto-generated captions, and their translations. The most prevalent category was auto-generated English captions, which accounted for 54.3% of the dataset ($n = 9,464$). An additional 25.4% ($n = 4,430$) of videos had translated auto-generated captions, further increasing accessibility for non-English content.

Only a small proportion of videos included manually created captions, with 898 videos (5.2%) containing original English captions uploaded by creators. Another 141 videos (0.8%) featured English transcripts translated from manually created captions in other languages.

We also reported engagement figures to illustrate the differences brought about by the availability of different types of captions. These are reported using median values rather than means to better reflect central tendencies in the presence of right-skewed distributions. View, like, and comment counts in the datasets follow a long-tailed distribution, where a small fraction of viral videos accumulate disproportionately large numbers, inflating the mean and obscuring the behavior of the majority. The use of medians thus offers a more robust summary of typical audience interaction.

While manually created and translated-from-manual transcripts are relatively rare, they have higher median engagement metrics. For instance, videos with manually created transcripts had a median view count of $15,323$, compared to just $1,655.5$ for those with auto-generated captions. They also received more comments (median = 6 vs. 1) and likes (median = 118 vs. 24).

Transcript Type	**Count**	**Share of All Videos**	**Median View**	**Median Comment**	**Median Like**
Manually Created	898	5.2%	**15,323**	6	118
Translated from Manually Created Captions	141	0.8%	11,482	21	165
Auto-Generated	**9,464**	**54.3%**	1,655.5	1	24
Translated from Auto-Generated Captions	4,430	25.4%	2,262.5	3	40.5
Total w/ Transcript	**14,933**	**85.6%**	-	-	-

Table 3.9: Availability and types of English transcripts in the final dataset

Language Distribution

Given the global nature of YouTube and the diversity of its creator base, understanding the linguistic composition of the dataset provides insight into its accessibility and potential reach. Each video in the dataset includes two key language metadata fields provided by the YouTube API: `audio_language` and `textual_language`, corresponding to the original caption language and the language of textual metadata (e.g., title, description), respectively.

Audio Language. As shown in Table 3.10, 59.05% of the videos ($n = 10,294$) had a specified audio language. Among those, various English dialects collectively dominated the corpus. The most frequent audio language tag was "English" (as labeled by YouTube), which alone accounted for 27.2% of all videos. When combined with other regional variants explicitly tagged by YouTube, such as English (United States), English (India), and English (United Kingdom), English language content collectively represented 43.9% of the dataset. Other widely represented languages included Hindi (4.72%) and Arabic (1.04%), and several regional Indian languages such as Telugu, Urdu, and Tamil.

Some languages with smaller representation (e.g., Telugu, Tamil) had relatively high median engagement: videos in Telugu, for example, had the highest median view count ($6,262$) and like count (141).

Top 10 Audio Languages	Count	Share of All Videos	Median View	Median Comment	Median Like
English	**4,748**	**27.23%**	3,714	3	51
English (United States)	1,224	7.02%	1,753	1	23
English (India)	1,028	5.90%	1,951.5	2	37
Hindi	823	4.72%	3,892	5	73
English (United Kingdom)	656	3.76%	3,959	3	61.5
Arabic	182	1.04%	5,976.5	3	100.5
Telugu	169	0.97%	**6,262**	6	**141**
Urdu	146	0.84%	959.5	4	27.5
Tamil	126	0.72%	4,839.5	**10**	109
Indonesian	105	0.60%	2,083	3	42
Other	1,087	6.23%	3,743	4	62
Total w/ Audio Language Specified	**10,294**	**59.05%**	-	-	-

Table 3.10: Top 10 most frequent languages by video audio metadata

Textual Language. Textual metadata (titles, descriptions) displayed more sparsity as shown in Table 3.11: only 21.74% of videos ($n = 3,790$) had a `textual_language` tag. English and its regional variants again dominated, with plain "English" tagged in 12.36% of all videos, followed by English (India), English (US), and English (UK). Other languages like Portuguese, Hindi, Arabic, and Spanish (Latin America) occurred less frequently, though several exhibited high median engagement levels. For instance, Spanish (LATAM) videos showed the highest median view ($15,716$) and like counts (511) among all tagged textual languages.

Top 10 Textual Languages	Count	Share of All Videos	Median View	Median Comment	Median Like
English	**2,155**	**12.36%**	2,989	2	44
English (India)	573	3.29%	753	2	20
English (United States)	348	2.00%	2,055	3	45
English (United Kingdom)	228	1.31%	3,581	3	53
Portuguese	37	0.21%	9,946	12	388
Hindi	34	0.20%	15,230	19	264.5
Arabic	30	0.17%	3,041	2.5	65.5
Spanish (Latin America)	27	0.15%	**15,716**	**27**	**511**
Indonesian	27	0.15%	3,394	4	40
German	24	0.14%	818.5	0	10.5
Other	307	1.76%	4,205	6	64
Total w/ Textual Language Specified	**3790**	**21.74%**	-	-	-

Table 3.11: Top 10 most frequent languages by video textual metadata

This discrepancy between audio and textual tagging coverage indicates possible limitations in metadata quality and creator-side language settings. Moreover, the sparse labeling on textual language suggests that many videos lack explicitly set metadata, relying instead on default or inferred settings.

Geographical Distribution

To understand the global provenance of educational content in data systems, we analyzed the `country` field associated with each video's originating channel. This metadata field, provided by the YouTube API, indicates the registered country of the channel creator and thus serves as a proxy for the geographic origin of production. Of the $7,587$ unique channels in the final dataset, $5,038$ (66.4%) provided a non-null country identifier, enabling a fairly comprehensive geographic analysis. These channels account for $13,213$ videos, approximately 75.8% of all content in the final dataset, making country-based patterns representative of broader production trends.

Table 3.12 summarizes the top 10 countries by channel count. The largest share of content in our dataset originates from India, which accounts for $2,100$ channels (41.7% of those with country registered), and $6,507$ videos (37.3% of all videos). Indian creators tend to operate mid-sized channels (median $2,240$ subscribers) with a median video output of 188 videos per channel. Their videos are moderately engaging, with a median of $3,886$ views, 59 likes, and 4 comments.

The United States, as the second-largest producer, contributed $1,051$ channels and $2,918$ videos

(16.7%). These channels typically show higher production activity (median 250 videos/channel) and more subscribers (median $5,330$), though video-level engagement is slightly lower than India's (median $2,763$ views, 36 likes, 2 comments).

Despite a smaller number of channels, Brazil exhibits exceptional performance across multiple indicators. It has the highest median number of videos per channel (265.5), the highest median subscriber count $(9,045)$, and the strongest per-video engagement: median of $6,590$ views, 235 likes, and 11 comments.

Although the United Kingdom only accounts for 2.2% of all videos and has 167 associated channels, it demonstrates relatively strong audience engagement. Its videos show a higher median view count $(6,385.5)$ and like count (61.5) than most other countries with more contributions. This suggests that while UK-based content is less prevalent in volume, it may attract more attention per video.

Country	Channels	Videos (%)	Median Video	Median Subs	Median View	Median Comment	Median Like
India	**2,100**	**6,507 (37.32%)**	188	2,240	3,886	4	59
United States	1,051	2,918 (16.74%)	250	5,330	2,763	2	36
Pakistan	249	566 (3.25%)	155	1,380	922.5	1	18
United Kingdom	167	384 (2.20%)	175	5,570	6,385.5	5	61.5
Canada	114	227 (1.30%)	138	2,415	2,633	3	35
Indonesia	114	163 (0.93%)	89	1,020	1,360	1	25
Germany	92	207 (1.19%)	191	2,785	3,142	4	55
Brazil	72	129 (0.74%)	**265.5**	**9,045**	**6,590**	**11**	**235**
Egypt	57	123 (0.76%)	197	4,390	1,487	2	32.5
Australia	48	143 (0.82%)	147.5	4,790	3,844	1	73
Other	974	1,846 (10.59%)	154.25	2,885	2,670.5	2	32.5
Total w/ Country Specified	**5,038**	**13,213(75.79%)**	-	-	-	-	-

Table 3.12: Top 10 countries by channel count and associated median metrics. ”Median Video” and ”Median Subs” are calculated per channel; ”Median View”, ”Median Like”, and ”Median Comment” are calculated per video.

Topic Distribution

To examine how comprehensively the collected dataset reflects the curriculum-informed subtopic space, we analyzed the topical coverage of videos based on their matched search queries. Each video in the dataset was originally retrieved via one or more queries corresponding to the 38 subtopics proposed in the educator survey by Miedema et al. [58]. However, due to the granularity and fragmentation of the original subtopic list, we consolidated several closely related subtopics into higher-level categories to facilitate summarization and pattern recognition.

For example, the subtopic “relational theory: relations, tuples and attributes” was aggregated under the broader topic “relational theory”, and three separate subtopics on ”database optimization: indexing”, ”database optimization: query optimization”, and ”database optimization: execution plans” were unified under “database optimization” based on the heading topic. This consolidation enabled us to visualize broader trends in topic prevalence while still retaining the alignment with the original curriculum structure.

Using the query-to-topic mapping, we computed the number of videos associated with each topic. The resulting topics and their distribution are shown in Table 3.13. It presents the number of videos matched to each topic (based on query associations), alongside median values for view count, comment count, like count, and video duration. The most prominent topic by volume is SQL, accounting for 27.3% of all videos $(n = 4,915)$, followed by Database Normalization (10.2%) and Data Mining (9.7%).

However, higher volume does not always equate to higher viewer engagement. For instance, videos under Data Visualization, while comprising only 2.6% of the dataset, achieved the highest median like count (175), and are among the top three in comment volume (median = 5). Data Security and Access Management also stands out, with the highest median view count $(12,716)$ and strong engagement across other metrics, despite representing under 2% of videos. Additionally, Distributed Database Management Systems and Object-Oriented Data Models, while not large in volume, demonstrated relatively higher overall engagement levels.

At the other end of the distribution, topics such as Semi-Structured Traditional Data Models (0.16%) and Relational Theory (0.79%) are the most underrepresented, with low audience engagement. Notably, Semi-Structured Data Models also exhibited the longest median video duration $(23.13$ minutes).

Topics like Database Scalability, Tuple Relational Calculus, and Data Privacy and Ethics exhibited both relatively low volume and the lowest median engagement.

Topic	Videos (%)	Median View	Median Comment	Median Like	Median Duration (min)
SQL	**4,915 (27.28%)**	3,106	3	51	8.10
Database Normalization	1,844 (10.23%)	1,153.5	2	25	11.38
Data Mining	1,752 (9.72%)	2,377.5	2	28	9.75
Database Optimization	1,164 (6.46%)	883.5	1	14	10.92
Data Modeling	979 (5.43%)	5,382	4	63	13.68
Concurrency Control and Isolation Levels	777 (4.31%)	1,312	1	25	13.90
Database Back-ups and Recovery	697 (3.87%)	4,150	2	36	10.45
Logical and Physical Data Independence	576 (3.20%)	1,152.5	2	21	10.51
Data Warehousing	504 (2.80%)	2,420.5	3	47.5	8.88
Functions and Stored Procedures	500 (2.77%)	2,031	4	29	11.92
Relational Algebra	491 (2.72%)	5,582	3	89	12.63
Database Scalability	477 (2.65%)	388	0	6	14.73
Data Visualization	467 (2.59%)	10,212	**5**	**175**	12.33
NoSQL Database Management Systems	453 (2.52%)	512	1	12	6.20
Tuple Relational Calculus	405 (2.25%)	402	0	8	11.48
Distributed Database Management Systems	380 (2.11%)	7,209.8	**5**	80	9.18
Database Management Systems Components	357 (1.98%)	2,269	4	38	8.82
Data Security and Database Access Management	347 (1.93%)	**12,716**	4	124	11.57
Transaction Processing	298 (1.65%)	3,457	3	44	12.09
Data Privacy and Ethics	240 (1.33%)	507.5	1	5.5	11.30
Object-Oriented Data Models	226 (1.25%)	9,664	**5**	100.5	10.79
Relational Theory	142 (0.79%)	1,011	2	24.5	13.68
Semi-Structured Traditional Data Models	28 (0.16%)	3,083.5	0	19	**23.13**

Table 3.13: Data systems topics by video count and median engagement metrics.

To further explore topic-specific trends over time, Figure 3.4 plots the annual count of videos published for each consolidated topic from 2005 to 2024. This faceted line chart highlights how the presence of different data systems topics has evolved on YouTube, both in timing and intensity. Overall, most topics show a clear upward trend after 2020, though the rate and consistency of growth differ significantly across categories. The most prominent growth is seen in SQL, which experienced an explosive rise in 2020, reaching nearly $1,000$ videos in 2023, far exceeding any other topic. Topics like Database Normalization and Data Mining also saw notable spikes during 2019–2021, though their growth slowed or declined in the subsequent years. Other foundational topics, such as Database Optimization, Data Modeling, Data Warehousing, Database Scalability, and Data Visualization, exhibited more steady, incremental growth, especially Data Visualization, which saw its video count rise consistently post-2020. Meanwhile, niche or more theoretical topics such as Object-Oriented Data Models, Relational Theory, and Semi-Structured Data Models remained a steady presence but relatively underrepresented over time.

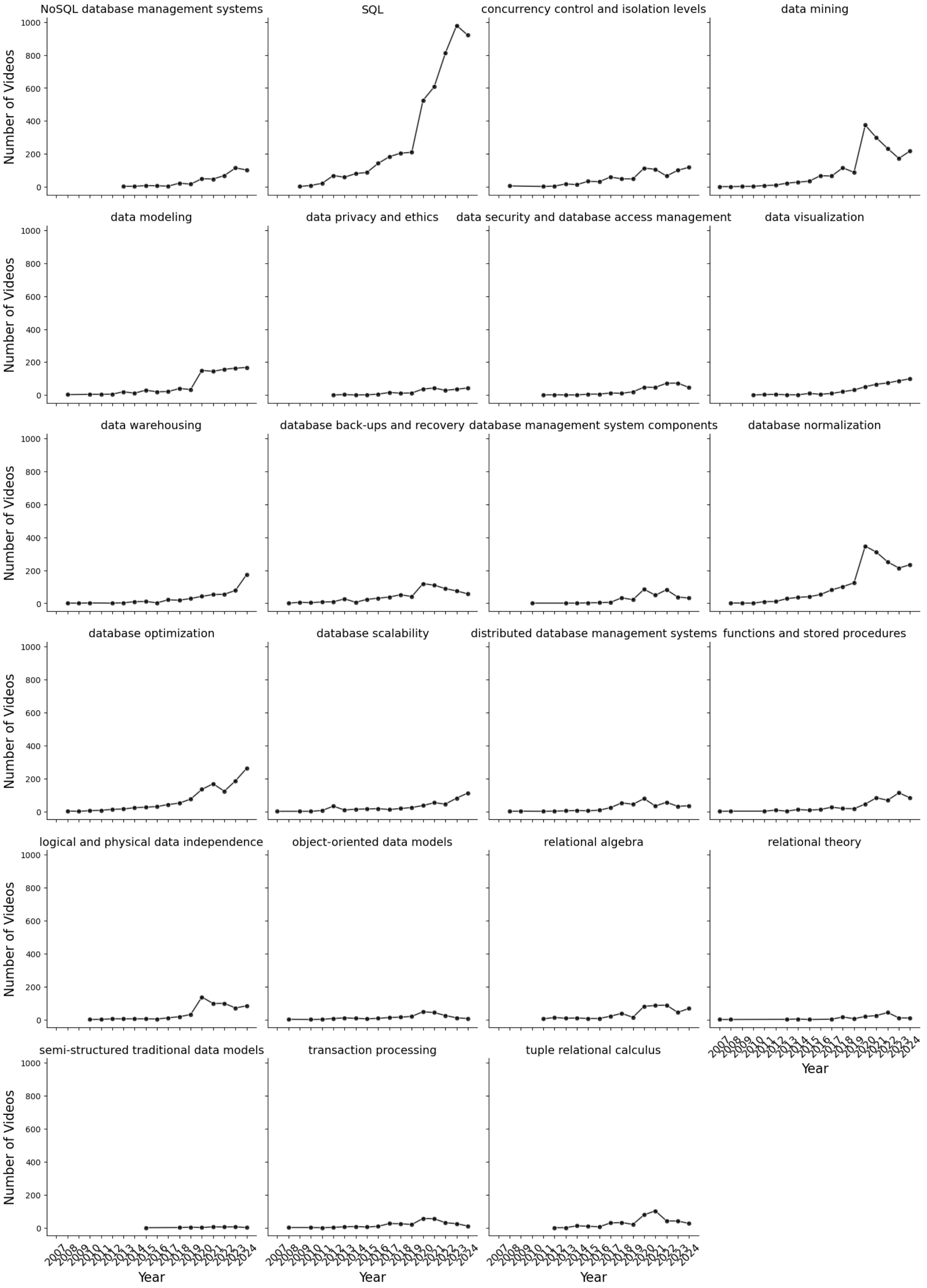


Figure 3.4: Temporal distribution of video production across 23 data systems topics (2007–2024)

4
Audience Engagement Modeling

This chapter presents the analytical framework and empirical findings for modeling audience engagement with educational YouTube videos on data systems. Section 4.1 describes the analytical procedures used to address RQ3, including the construction of an engagement index, the engineering of structural, linguistic, and contextual features, and the implementation of correlation and regression modeling. Section 4.2 presents the results of the statistical modeling, identifying which features are significantly associated with audience engagement.

4.1. Methods

To address *RQ3: Which features of educational YouTube videos on data systems are associated with audience engagement?*, we conducted a statistical modeling analysis linking video, channel, content, and language level attributes to audience responses. The modeling goal was to identify both the native and derived features affecting audience engagement of data systems educational videos on YouTube, measured through views, likes, and comments. This section describes the construction of the engagement metric, feature engineering, correlation analysis, and modeling techniques employed.

4.1.1. Feature Engineering: Engagement Metric Construction

To model audience engagement with educational YouTube videos on data systems, three behavioral indicators were selected: views per day, likes per day, and comments per day. These metrics reflect the frequency and intensity of user interaction with video content and collectively represent key dimensions of engagement: exposure, positive evaluation, and participatory response.

While past research has often combined engagement as a single composite score, typically calculated as a weighted sum of likes, comments, shares, and other interaction metrics [33, 41, 43, 91], such approaches can mask variation across different types of user responses. A study [82] found that image-based posts on social media tend to attract more likes than comments, while text-based posts elicit more comments than likes, suggesting that different content characteristics may drive different types of engagement. To address such concerns, Fischer et al. applied principal component analysis (PCA) to four engagement metrics obtained from the YouTube API (views, likes, dislikes, comments), extracting two distinct latent dimensions: popularity and polarity [24]. Inspired by this approach, we adopted a similar dimensionality reduction strategy, adapted to the specific constraints of our dataset. By the time our study was conducted, YouTube API access policies had changed, and interaction data on dislikes and shares were no longer publicly available. Consequently, our thesis focused on the three remaining accessible metrics: views (number of times a video was played), likes (number of times viewers clicked the ”like” button), and comments (number of viewers left a comment). To normalize for video lifespan, all metrics were converted to per-day rates: views/day, likes/day, and comments/day.

Descriptive analysis performed before in subsection 3.2.2 and Shapiro-Wilk tests revealed that all three engagement metrics were non-normally distributed ($p < .001$), exhibiting the strong skewness typical of social media interaction data. Accordingly, Spearman’s rank correlation was used to assess their relationships. The results showed high correlations between these metrics ($\rho = 0.93$ between views and likes, $\rho = 0.83$ between views and comments, and $\rho = 0.86$ between likes and comments). These high correlations suggested substantial redundancy, making dimensionality reduction appropriate. To construct a composite measure while preserving shared variance, we applied Principal Component Analysis (PCA) to the three engagement metrics.

As shown in the scree plot in Figure 4.1, the first principal component accounted for 82.2% of the total variance, capturing the common engagement signal across views, likes, and comments. Figure 4.1 presents the loadings of each original variable on the first principal component. All three metrics contributed substantially to the component, supporting its interpretation as a general engagement index. This component, referred to as the *Engagement Index*, was retained as the measure of audience engagement in the subsequent regression modeling.

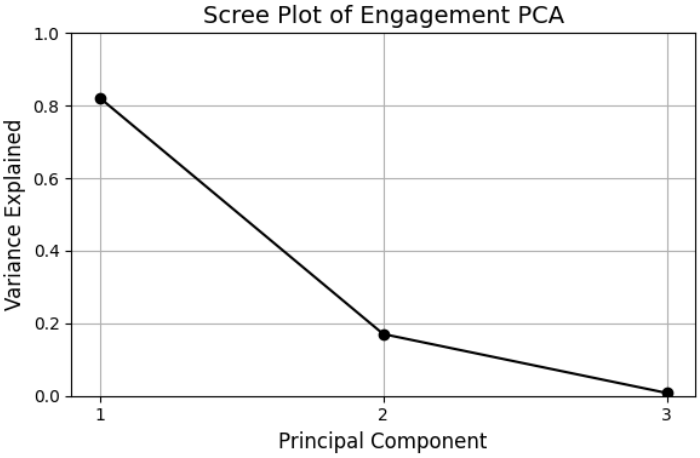


Figure 4.1: Scree plot of engagement PCA showing the variance explained by each principal component.

Variable	**PC1 Loading**
Views per day	0.540
Likes per day	0.632
Comments per day	0.555

Table 4.1: Loadings of engagement variables on the first principal component (PC1).

Given that the engagement index exhibited a highly skewed distribution with a long right tail, we applied a log1p transformation to the engagement index before regression modeling. All subsequent analyses, including feature screening and regression modeling, used the log1p-transformed engagement index as the dependent variable.

4.1.2. Feature Engineering: Covariates Extraction

To identify factors associated with audience engagement, we extracted a comprehensive set of features encompassing structural, linguistic, affective, and contextual attributes. These features were grouped into three major types: interval, binary, and nominal variables. Informed by prior research on science communication and video popularity on YouTube [95, 24], our selection included both directly available metadata from the dataset and derived variables constructed through computational methods.

Interval Variables

The set of interval-scale variables includes both native features from the dataset and variables that were derived through feature transformation or text analysis. These variables are:

1. video duration: length of the video in seconds.
2. subscriber count: number of subscribers to the video's hosting channel.
3. channel productivity: a derived metric indicating how frequently a channel publishes content. It has been shown to be associated with video popularity in prior studies of science content on YouTube [95]. It was calculated as the ratio of the total number of videos on a channel to the number of days since the channel was created.
4. number of tags: number of tags manually created by the video uploader to describe or categorize the video content.

5. channel-level view count: total number of views accumulated by the channel up to the time of data collection.
6. video age: number of days since the video was published.
7. channel age: number of days since the channel was first created.
8. title length: number of words in the video title, used as a proxy for headline complexity.
9. valence: a derived linguistic-affective feature representing the average emotional polarity of the words in a video's transcript. Following the method proposed by Fischer et al. [24]. Valence was calculated as the average difference between positive and negative scores across all valid sentiment-bearing words:

$$\text{valence} = \frac{1}{|\mathcal{W}|} \sum_{w_j \in \mathcal{W}} (\text{pos}(w_j) - \text{neg}(w_j)) \tag{4.1}$$

   where $\mathcal{W}$ denotes the set of words in the transcript that carry sentiment scores according to SentiWordNet, and pos(w_j), neg(w_j) refer to the positive and negative sentiment scores assigned to word w_j, respectively.

10. density: computed alongside valence using the same transcript-level analysis, density captures the proportion of words in the transcript that were identified as emotionally expressive (i.e., those with non-zero sentiment scores in `SentiWordNet`) [24]. It was calculated as:

$$\text{density} = \frac{1}{N} \sum_{j=1}^{N} \delta(w_j) \tag{4.2}$$

   where N is the total number of tokens in the transcript and $\delta(w_j)$ is an indicator function such that:

$$\delta(w_j) = \begin{cases} 1 & \text{if } w_j \in \mathcal{W} \\ 0 & \text{otherwise} \end{cases}$$

   Both valence and density were extracted using the following process: transcripts were cleaned and tokenized, stopwords were removed, tokens were part-of-speech (POS)-tagged and lemmatized. Each sentiment-carrying word was identified using its `WordNet` POS tag and scored using the first matching `SentiWordNet` synset. Words without sentiment scores or outside recognized POS categories were excluded. The resulting per-transcript values were computed using the custom functions applied across the full dataset.

11. readability score: a readability metric based on the Flesch-Kincaid Grade Level (FKGL), calculated using the `textstat.flesch_kincaid_grade` function. This score estimates the U.S. school grade level required to understand the video transcript, serving as a proxy for the linguistic accessibility of the educational content.

Among these, all text-based variables (i.e., valence, density, and readability score) were computed only for $1,039$ videos with creator-uploaded English captions (English translations included). After excluding entries with empty values for all interval variables, $1,002$ videos remained, ensuring sufficient input quality for analysis. This restriction was necessary due to the poor quality of many auto-generated captions on YouTube. In some cases, auto-captioning systems produced transcripts for videos that contained no speech at all, such as screen recordings or background music with no narration, resulting in meaningless or repetitive text. In other cases, even when speech was present, automatic recognition frequently misidentified words or omitted sentence boundaries, undermining the reliability of downstream linguistic analyses. By focusing only on human-curated transcripts, we aimed to ensure that valence, density, and readability scores were calculated on semantically meaningful content rather than noise or artifacts of automated captioning. Additionally, for the readability score, we observed that some transcripts, even when creator-uploaded, still lacked basic punctuation such as periods and commas, which are essential for accurate sentence boundary detection in Flesch–Kincaid computation. To address this issue, all transcripts were further processed using a multilingual punctuation restoration model `fullstop-punctuation-multilang-large` [27]. This model, applied after text cleaning (e.g., removal of bracketed metadata and whitespace normalization), restored punctuation marks (i.e., period, comma, equation mark, hyphen, colon) for likely sentence and clause boundaries to improve sentence segmentation quality. Only creator-uploaded transcripts that passed this preprocessing step were used to compute the Flesch–Kincaid readability grade, ensuring the resulting scores reflected linguistically coherent and punctuated input text.

This derivation process of valence, density, and readability enabled the modeling of latent or abstract video qualities that are not explicitly captured in platform metadata but may nevertheless influence user engagement.

Binary Variables

Two binary variables were initially considered as potential covariates of audience engagement:

1. definition: indicating whether the video was uploaded in high-definition (HD) format.
2. paid product placement: indicating whether the video was flagged as containing paid product placement or sponsorship.

Both variables are directly obtainable in the YouTube metadata. In theory, HD quality could be associated with more professional production standards and better viewer experience, while product placement might signal commercial intent, potentially influencing audience perception and interaction.

Nominal Variables

Several nominal variables were extracted to capture linguistic and contextual attributes that could influence audience engagement:

1. audio language: the language of the video's default audio track.
2. textual language: the language of the text in the video's title and description.
3. country: the country associated with the channel.
4. topic: assigned based on search terms following the same categories as in Table 3.13.

Together, the feature engineering process produced a set of covariates that combined platform accessible metadata with computationally derived variables from transcript text. This multi-dimensional feature set allowed for the modeling of both structural properties (e.g., channel and video structural characteristics) and latent qualities (e.g., affective tone, linguistic complexity) of educational videos.

4.1.3. Correlation Analysis

Following feature engineering, we performed correlation analysis to examine the relationships between candidate covariates and the dependent variable, log1p-transformed engagement index. This stage aimed to provide an overview of how individual features relate to engagement and to inform the subsequent regression modeling. The process combined distributional assessment, visual inspection of variable relationships, and calculation of correlation coefficients.

Interval Variable

Histograms were generated for each interval variable to assess distributional characteristics as shown in Figure 4.2. Several variables, including subscriber count, channel-level view count, and video duration, exhibited extreme positive skewness with long right tails as indicated in subsection 3.2.2. To address this, we applied a log1p transformation to these variables, compressing their range and reducing the influence of outliers while preserving relationships. Variables that were already symmetrically distributed, for which log transformation was inappropriate (e.g., valence, density), or distributions after log1p transformation that remained concentrated in lower ranges (i.e., channel productivity) were retained in their original scale.

Scatterplots were constructed to visualize the relationships between each covariate and the log1p-transformed engagement index as indicated in Figure 4.3. This step enabled the detection of possible trends and potential high-leverage points that could influence correlation estimates or model stability. The scatterplots confirmed that variables with log1p transformations exhibited more balanced distributions along both axes.

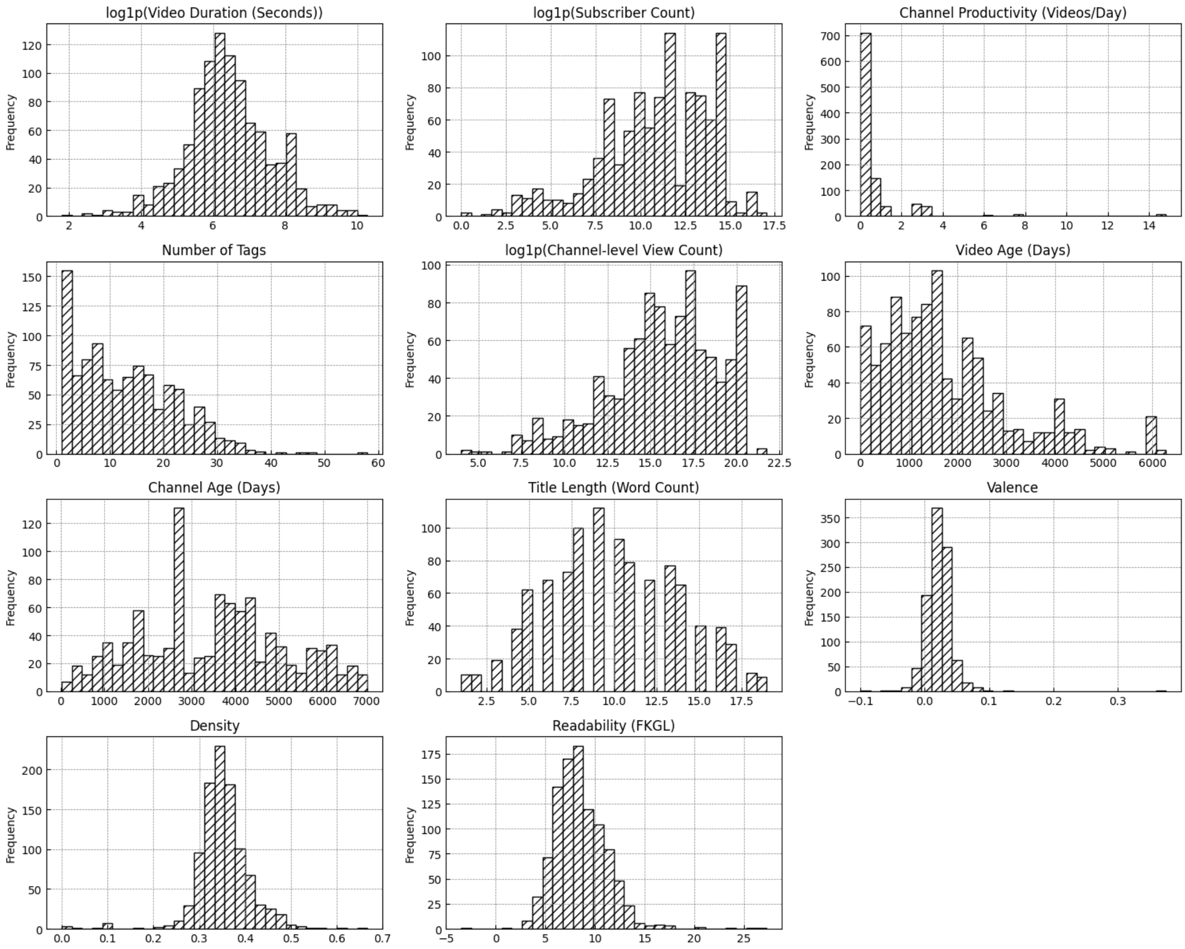


Figure 4.2: Distributions of interval covariates, with log1p transformation applied where appropriate.

Spearman's rank correlation coefficients were calculated to assess the correlations between each covariate and the log1p-transformed engagement index. The results are summarized in Table 4.2. The strongest positive associations were observed for *log1p(Subscriber Count)* ($\rho = 0.56$, $p < 0.001$) and *log1p(Channel-level View Count)* ($\rho = 0.50$, $p < 0.001$). *Readability (FKGL)* showed a moderate negative association ($\rho = -0.31$, $p < 0.001$). Variables such as *Number of Tags*, *log1p(Video Duration)*, *Title Length*, and *Channel Productivity* had weaker positive correlations (all $\rho > 0.3$). *Density*, *Video Age*, *Channel Age*, and *Valence* exhibited weak associations (all $|\rho| < 0.11$).

Binary Variables

Point-biserial correlation coefficients were computed to assess the relationships between the binary covariates and the log1p-transformed engagement index. The results showed weak positive associations: $\rho = 0.123$ ($p = 8.91 \times 10^{-5}$) for *High-definition (HD)* and $\rho = 0.137$ ($p = 1.35 \times 10^{-5}$) for *Paid Product Placement*.

However, initial inspection of their distributions revealed extreme class imbalance: 93% of the videos were published in HD, and 99.4% of the videos did not contain any declared product placement. Because of this imbalance, these variables offered minimal variance and informational value for explaining differences in engagement. Specifically, HD video is now a normal form for most videos, so its research value as a variable for distinguishing between different levels of engagement is limited. Similarly, paid product placement is self-reported by uploaders, and the proportion of videos that declared such placement is less than 1%, so the information is both unreliable and of less meaning for analysis. For these reasons, both variables were excluded from the following regression analysis.

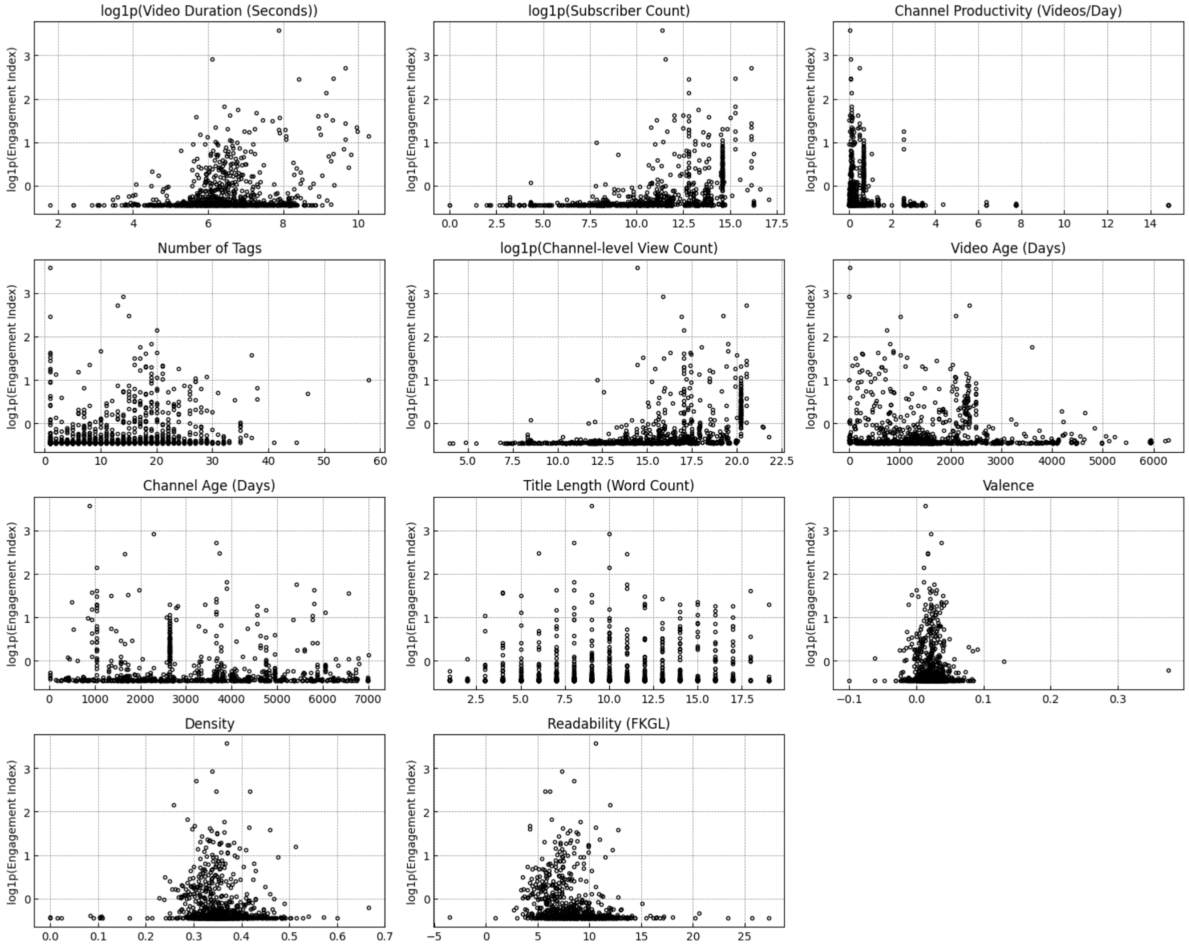


Figure 4.3: Scatterplots of interval covariates versus log1p-transformed engagement index.

Nominal Variables

For nominal covariates, associations with the log1p-transformed engagement index were evaluated using different correlation measures based on the structure of the nominal variables. For categorical variables with multiple levels (audio language, textual language, and country), the Kruskal-Wallis H-test was used to assess whether engagement index distributions differed significantly across categories. For topic indicators, which were coded as binary presence variables (1 = present, 0 = absent) as a video can be assigned to multiple topics, point-biserial correlations were computed to quantify the association between topic presence and engagement.

The Kruskal-Wallis tests indicated significant differences in engagement index distributions across categories of audio language ($H = 59.58$, $p < 0.001$), textual language ($H = 33.98$, $p < 0.001$), and country ($H = 24.47$, $p < 0.001$) as in Table 4.3. Boxplots in Figure 4.4 and Figure 4.5 illustrate these distributions for languages and countries with over 30 samples, showing that videos in English (India) and channels from India tended to have higher engagement indices, although videos from the UK have slightly higher median engagement.

For topic indicators, point-biserial correlations revealed generally weak associations with engagement index as seen in Table 4.4. The strongest correlation was observed for *SQL* ($\rho = 0.18$, $p < 0.001$), with other topics showing negligible or no association. Figure 4.6 provides a visual summary of engagement index distributions by topic presence.

Therefore, topic indicators were excluded from the regression model as correlation analysis revealed generally weak associations with the engagement index (all $|\rho| < 0.10$ except for SQL). For the simplicity of the model, only the SQL topic indicator was retained for further topic-wise analysis.

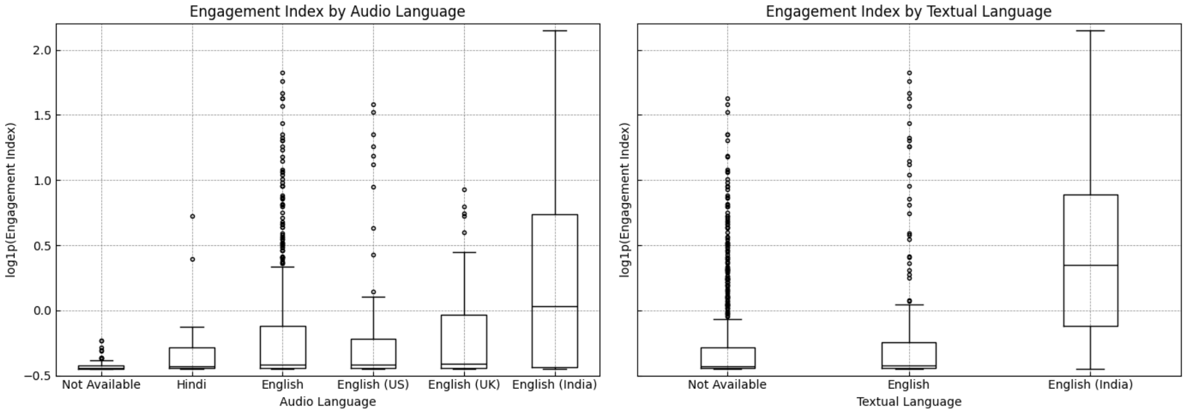


Figure 4.4: Boxplots of log1p-transformed engagement index by audio language (left) and textual language (right), including only languages with over 30 samples.

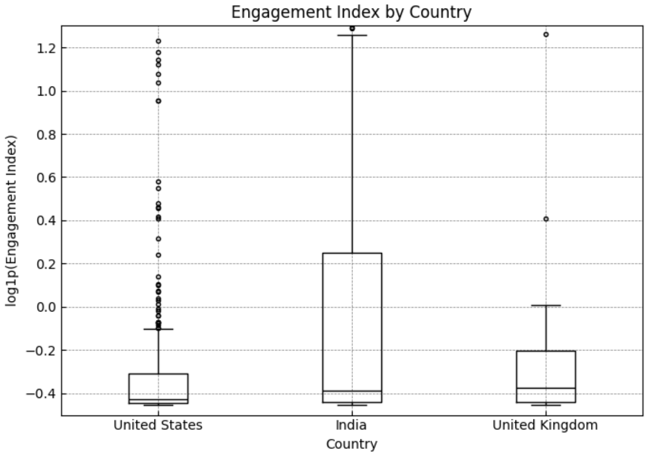


Figure 4.5: Boxplot of log1p-transformed engagement index by country, including only countries with over 30 samples.

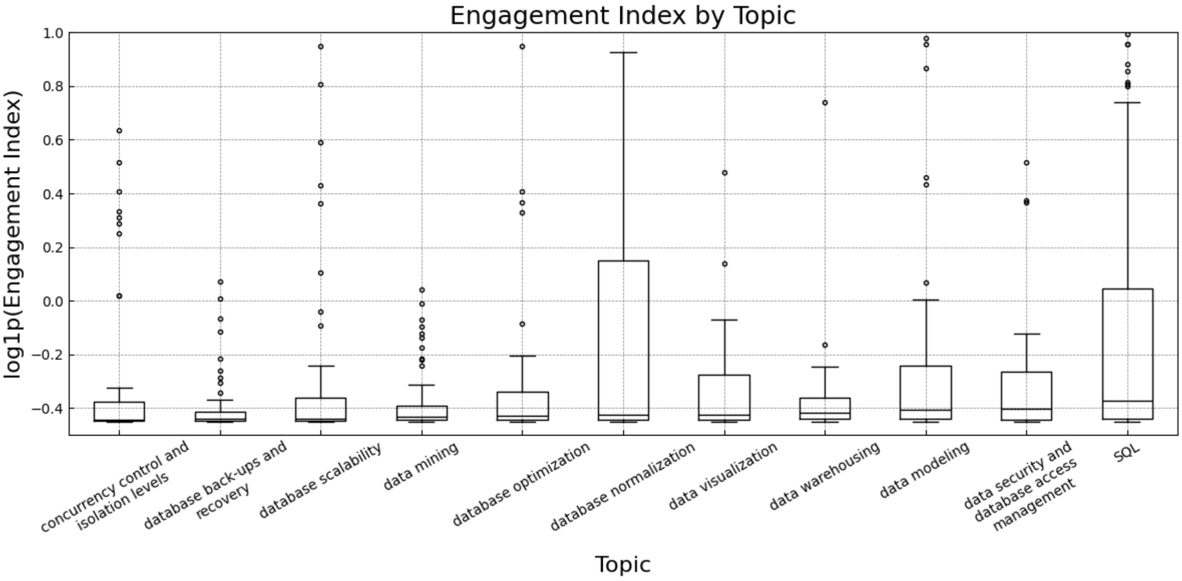


Figure 4.6: Boxplots of log1p-transformed engagement index by topic presence.

Variable	Spearman ρ	p-value
log1p(Subscriber Count)	0.563	5.73×10^{-85}
log1p(Channel-level View Count)	0.504	1.00×10^{-65}
Readability (FKGL)	-0.312	3.97×10^{-24}
Number of Tags	0.241	1.12×10^{-14}
log1p(Video Duration (Seconds))	0.173	3.62×10^{-8}
Title Length (Word Count)	0.150	1.82×10^{-6}
Channel Productivity (Videos/Day)	0.117	2.00×10^{-4}
Density	-0.109	5.33×10^{-4}
Video Age (Days)	-0.088	5.23×10^{-3}
Channel Age (Days)	-0.075	1.74×10^{-2}
Valence	-0.052	9.95×10^{-2}

Table 4.2: Spearman's rank correlations between interval covariates and log1p-transformed engagement index.

Variable	H-statistic	p-value
Audio Language	59.58	1.49×10^{-11}
Textual Language	33.98	4.19×10^{-8}
Country	24.47	4.87×10^{-6}

Table 4.3: Kruskal–Wallis test results for language and country covariates.

Topic	ρ	p-value
SQL	0.183	5.62×10^{-9}
Data Mining	-0.050	0.11
Data Modeling	0.037	0.24
Database Normalization	0.050	0.12
Database Optimization	-0.040	0.21
Database Back-ups and Recovery	-0.080	0.01
Data Visualization	0.043	0.18
Concurrency Control and Isolation Levels	-0.026	0.41
Database Scalability	-0.021	0.50
Data Security and Database Access Management	-0.010	0.74
Data Warehousing	-0.016	0.62

Table 4.4: Point-biserial correlations between topic presence and log1p-transformed engagement index.

4.1.4. Modeling Procedures

An ordinary least squares (OLS) regression model was used to examine the relationships between covariates and the log1p-transformed engagement index. All interval variables were included as covariates, reflecting structural, temporal, and content characteristics of the videos and channels. The SQL topic indicator was included as the sole topic variable, as prior correlation analysis had shown negligible associations for other topics. Nominal variables (audio language, textual language, and country) were dummy-coded after merging rare categories (fewer than 30 observations) and data with unspecified nominal variables into an 'Other' group to ensure model simplicity and interpretability. All dummy-coded variables were processed using one-hot encoding, with one reference category omitted to avoid multicollinearity due to the dummy variable trap. The final model was estimated on $1{,}002$ videos with creator-uploaded captions and complete data for the selected interval features. We required complete data for all interval variables because these represent core structural features of the videos or channels that are generally available and reliable, apart from rare instances where the YouTube API failed to return valid values due to technical issues. These cases were excluded to preserve data integrity. In contrast, nominal variables such as language and country are primarily based on creator self-reports or optional metadata and exhibit substantial missingness. To maximize sample retention and reduce bias from listwise deletion, records with missing nominal data were assigned to the 'Other' category along with rare groups.

After building a multiple linear regression model, variance inflation factors (VIFs) were computed to assess multicollinearity among covariates. All analyses were conducted using the `statsmodels` package in Python, and model summaries reported coefficients, standard errors, t-statistics, p-values, R^2, and adjusted R^2.

4.2. RQ3 Results: Factors Associated with Audience Engagement

The results of the multiple linear regression model explained approximately 35% of the variance in the log1p-transformed engagement index ($R^2 = 0.351$, adjusted $R^2 = 0.335$). The overall model was statistically significant ($F(24, 977) = 22.01$, $p < 0.001$). Table 4.5 summarizes the regression coefficients, standard errors, p-values, and 95% confidence intervals for all covariates in the model. Several co-

Covariate	Coef	Std Err	t	P>|t|	Confidence Interval (95%)
Intercept	−1.569	0.186	−8.457	0.000	[−1.933, −1.205]
Video Duration (log1p)	0.078	0.012	6.717	0.000	[0.055, 0.101]
Subscriber Count (log1p)	0.040	0.018	2.180	0.029	[0.004, 0.076]
Channel-level View Count (log1p)	0.036	0.018	1.980	0.048	[0.000, 0.071]
Channel Productivity	−0.056	0.009	−5.936	0.000	[−0.075, −0.038]
Number of Tags	0.001	0.002	0.668	0.504	[−0.002, 0.004]
Video Age (Days)	-2.05×10^{-5}	1.31×10^{-5}	−1.564	0.118	$[-4.62 \times 10^{-5}, 5.22 \times 10^{-6}]$
Channel Age (Days)	-5.84×10^{-5}	9.99×10^{-6}	−5.849	0.000	$[-7.8 \times 10^{-5}, -3.88 \times 10^{-5}]$
Title Length (Word Count)	−0.0075	0.004	−1.975	0.049	[−0.015, −0.00005]
Valence	−0.305	0.589	−0.517	0.605	[−1.460, 0.851]
Density	0.404	0.249	1.624	0.105	[−0.084, 0.893]
Readability (FKGL)	−0.0096	0.005	−1.770	0.077	[−0.020, 0.001]
SQL Topic	0.117	0.031	3.817	0.000	[0.057, 0.177]
Audio: Other	−0.095	0.069	−1.375	0.170	[−0.231, 0.041]
Audio: English	0.059	0.052	1.132	0.258	[−0.043, 0.162]
Audio: English (UK)	0.027	0.080	0.339	0.734	[−0.130, 0.184]
Audio: English (India)	0.028	0.110	0.254	0.800	[−0.189, 0.245]
Audio: English (US)	0.113	0.068	1.664	0.097	[−0.020, 0.246]
Audio: Hindi	−0.108	0.092	−1.180	0.238	[−0.289, 0.072]
Text: Other	0.014	0.056	0.255	0.799	[−0.096, 0.125]
Text: English	0.012	0.033	0.371	0.711	[−0.052, 0.076]
Text: English (India)	0.377	0.108	3.481	0.001	[0.164, 0.590]
Country: India	0.059	0.075	0.788	0.431	[−0.088, 0.206]
Country: Other	0.116	0.074	1.584	0.114	[−0.028, 0.261]
Country: US	0.009	0.074	0.114	0.909	[−0.137, 0.154]

Table 4.5: OLS regression results for covariates of log1p-transformed engagement index.

variates were significantly associated with the engagement index. Among the interval variables, *log1p duration* ($\beta = 0.078$, $p < 0.001$), *log1p subscriber count* ($\beta = 0.040$, $p = 0.029$), and *log1p channel view count* ($\beta = 0.036$, $p = 0.048$) showed positive associations with engagement. *Channel productivity* exhibited a negative association ($\beta = -0.056$, $p < 0.001$), as did *channel age* ($\beta = -5.84 \times 10^{-5}$, $p < 0.001$). *Title length* showed a marginal negative association ($\beta = -0.0075$, $p = 0.049$).

For nominal variables, the presence of the SQL topic ($\beta = 0.117$, $p < 0.001$) was associated with higher engagement. Among the language variables, videos with *textual language: English (India)* were positively associated with engagement ($\beta = 0.377$, $p = 0.001$). Other language and country variables showed no statistically significant associations at the 5% level.

Covariate	VIF
Intercept	226.03
Video Duration (log1p)	1.22
Subscriber Count (log1p)	20.14
Channel-level View Count (log1p)	21.37
Channel Productivity	1.46
Number of Tags	1.38
Video Age (Days)	1.93
Channel Age (Days)	1.79
Title Length (Word Count)	1.39
Valence	1.08
Density	1.37
Readability (FKGL)	1.37

Covariate	VIF
SQL Topic	1.33
Audio: Other	2.60
Audio: English	4.30
Audio: English (UK)	1.57
Audio: English (India)	3.14
Audio: English (US)	2.24
Audio: Hindi	1.61
Text: Other	1.25
Text: English	1.24
Text: English (India)	2.38
Country: India	7.86
Country: Other	7.74
Country: US	8.01

Table 4.6: Variance inflation factors (VIFs) for covariates in the final model.

VIFs were examined to assess multicollinearity. Table 4.6 presents the VIF values for all covariates. Most of them had VIF values below 5, indicating no serious multicollinearity. As expected, country dummy variables exhibited higher VIFs (approximately $7-8$) might be due to their mutual exclusivity and overlap with language indicators. Log-transformed subscriber count and channel view count showed elevated VIFs (≈ 20), reflecting their correlation, but key coefficients remained interpretable and stable. The model's condition number was large (2.05×10^5), indicating potential numerical sensitivity, yet the results were consistent with theoretical expectations.

5

Textbook-Derived SQL Subtopics Coverage on YouTube

This chapter explores the extent to which YouTube videos cover SQL subtopics derived from standard database textbooks. Section 5.1 outlines our approach for identifying gaps and concentrations in topic coverage by aligning SQL-related YouTube content with a textbook-derived subtopic structure. This includes the construction of a SQL topic structure based on widely-used textbooks, the selection of a SQL-related video subset, and the evaluation of multiple classification strategies. Section 5.2 presents the findings from applying this classification to all SQL-related videos, revealing both heavily covered core areas of SQL and subtopics that are entirely absent or rarely exist in the YouTube corpus, thus highlighting both pedagogical strengths and blind spots in user-generated educational content on YouTube.

5.1. Methods

To address RQ4: *What gaps and popular areas exist in YouTube video coverage of data systems topics, based on alignment with academic textbook-derived topics?*, we conducted a focused subtopic coverage analysis using SQL as a case study. The choice of SQL was motivated by several reasons. (1) Prevalence: SQL emerged as the single most frequently covered topic in our YouTube dataset, accounting for approximately one-quarter of all curated data systems videos. (2) Engagement correlation: In the audience engagement modeling presented in chapter 4, the presence of SQL content in a video showed the strongest positive correlation with engagement metrics ($\rho \approx 0.18$), and SQL remained one of the most significant covariates of engagement in the final regression model, highlighting its particular resonance with viewers. (3) Curricular centrality: SQL is a foundational and well-structured topic in data systems education, taught in nearly all database courses and consistently covered across textbooks, which makes it ideal for deriving a textbook-aligned topic structure. (4) Scope feasibility: A comprehensive classification of all data systems topics at similar granularity would have resulted in a task beyond a manageable size, given the scale of the dataset and the number of subtopics involved. By focusing on SQL, we balanced between methodological rigor and feasibility, while still examining a topic that is representative and sufficiently rich to reflect how the topics derived according to data systems textbooks are represented in user-generated educational content on YouTube.

5.1.1. Textbook-Derived Subtopic Extraction

To analyze the instructional coverage of SQL topics across videos, we required a grounded and pedagogically relevant taxonomy of SQL subtopics. We sought to anchor our topic structure in established curricular sources. SQL subtopics covered in textbooks tend to reflect instructional focuses and consensus within the academic community. Therefore, we first constructed a list of SQL keyword terms grounded in three textbooks. To achieve this, index terms related to SQL were collected from three widely used data systems textbooks:

1. *Database Management Systems (3rd Edition)* by Ramakrishnan and Gehrke
2. *Database System Concepts (7th Edition)* by Silberschatz, Korth, and Sudarshan
3. *Database Systems: The Complete Book* by Garcia-Molina, Ullman, and Widom

Using these textbooks' indexes ensured coverage of the full breadth of SQL concepts as presented in formal curricula. From each textbook, all index terms along with their corresponding page numbers were collected. We then identified the core SQL chapters in each book: *Database Management Systems* Chapter 5: SQL: Queries, Constraints, Triggers, *Database System Concepts* Part 1: Relational Languages, and *Database Systems: The Complete Book* Chapter 6: The Database Language SQL, and filtered the index terms to include only those appearing within the page ranges of these chapters. The resulting raw keyword lists for each textbook are provided in Appendix B.1.

The combined raw list contained a total of 505 unique index terms, which included many synonyms, variations in phrasing, and some entries that were not directly related to SQL concepts (e.g., author names or marginal terms). This resulted in a long, unstructured list that was impractical as a basis for further analysis. To transform this into a usable structure for subtopic classification, we then used the OpenAI o3 model, guided by a structured prompt provided in Appendix A.2, to group the terms into higher-level subtopics that reflect coherent and meaningful SQL concepts. The proposed groupings were subsequently refined by removing redundant synonyms and unrelated marginal terms to streamline the keywords in each group and through domain expert review to ensure that the subtopics were sound in curricular and suitable for further analysis. This process resulted in a final structure comprising 70 subtopics, which are listed in Table 5.1, with their associated 406 unique textbook-derived keyword terms provided in full in Appendix B.2.

Active Databases	Aggregate Functions	Aliases & Correlation
Arity	Atomicity & Domains	Authorization & Privileges
Backup & Recovery	Business Logic	Cartesian & Product
Catalogs & Metadata	Change Tracking & Delta	Common Language Runtime (CLR)
Cursor Operations	Data Definition Language (DDL)	Data Manipulation Language (DML)
Data Types - Large Objects	Data Types - Scalar	Database Systems
Difference & EXCEPT	Dirty Data	Domain & Check Constraints
Duplicate Handling	Embedded & Dynamic SQL	Example Databases
Exceptions & Debugging	Expressions & Syntax	Fetch/Result APIs
Group BY & Having	Hierarchies	Identity Columns
Index	Integrity Constraints	Join Operations
Key Constraints	Language Integrated Query (LINQ)	Logical Connectives
Null & Unknown Handling	Operating Systems	Ordering & Limits
Partitioning	Pointers	Prepared Statements
Procedures & PSM	Programming Languages	Projection & Project Operation
Queries & Paradigms	Recursive Queries	Referential Integrity
Relational Model & Algebra	Row-Level Security	Scalar Functions
Schema	Security	Select Variants
Sequence	Set & Assignment	Set Operations
SQL Standards & History	Statistics	String Functions
Subqueries	Table	Table Functions
Temporal Concepts	Transactions & Isolation	Triggers
Type	View	Windowing & Pivoting
WITH Clauses		

Table 5.1: SQL subtopics (sorted alphabetically). The full mapping to textbook-derived keyword terms is provided in Appendix B.2.

5.1.2. SQL Video Subset Selection

From our dataset of $17,434$ curated educational videos on data systems, we extracted the subset of videos related to SQL for this SQL-dedicated analysis. Videos had been originally tagged by topic based on their search query matches (as did in subsection 3.2.2); we therefore filtered for all videos retrieved via SQL-related queries (e.g., “SQL”, “SQL select”, “SQL join”). This yielded $4,915$ videos identified as SQL-focused. Among these, $4,242$ videos included transcripts, either manually created or automatically generated. This subset formed the basis for subtopic classification. Videos without transcripts were excluded from this stage of analysis for the reason that our research question required examining the actual instructional content of each video, which cannot be reliably inferred without access to its actual content. While video titles and descriptions are available, titles are often too generic to identify which specific concepts are taught, and descriptions, when present, frequently consist of promotional or channel-related information rather than instructional substance. Therefore, transcripts were essential to enable meaningful and fine-grained classification of videos into SQL subtopics.

5.1.3. Video Classification Approach

We experimented with multiple methods to classify each video (specifically, each video’s transcript) into the 70 SQL subtopics defined above. Early attempts relied on keyword matching and embedding similarity, but these proved inadequate. In a first approach, we used `KeyBERT` [26], a BERT-based keyword extraction tool that identifies the most relevant words or phrases within a document by ranking terms based on their embedding similarity to the document as a whole. However, `KeyBERT` proved unsuitable for our task. The tool struggles when provided with a large candidate keyword set for direct matching. In our case, with 406 textbook-derived keyword terms across 70 subtopics, `KeyBERT` could not effectively rank the candidates in a way that reflected the true subtopic focus of the videos. Attempts to process candidates in batches, whether using subtopics directly or their associated keyword terms with post-hoc mapping to subtopics, failed to produce reliable rankings that meaningfully corresponded to the core educational focus of each video transcript. The same limitations applied when using `KeyBERT` purely for free-form keyword extraction without predefined candidates; the extracted keywords were often too generic or too fragmented to support accurate subtopic classification.

To further test whether this class of embedding-based similarity methods could be viable, we implemented a more general approach: representing both transcripts and subtopic (or their associated keyword terms) as high-dimensional vectors using sentence-transformer models and computing cosine similarity scores between them. We experimented with two variations: (1) directly comparing video transcript embeddings to subtopic embeddings, and (2) comparing transcript embeddings to the embeddings of individual keyword terms and aggregating or mapping the results back to subtopics. In both cases, although semantic matching allowed greater tolerance to lexical variation, the methods failed to reliably rank subtopics in a way that aligned with the true focus of the videos. Similarity scores often reflected spurious associations due to overlapping terminology in unrelated contexts, and did not provide a meaningful signal of topical focus. Another critical challenge in these embedding-based methods was the absence of a robust decision criterion for determining which subtopics to assign to a given video. Since videos could belong to zero, one, or multiple subtopics, naive strategies such as assigning the top-n most similar subtopics were unsatisfactory. We attempted to define a threshold by reviewing the literature. Rekabsaz et al. [71] proposed a general threshold for separating semantically related terms in word embeddings based on uncertainty analysis, but their method is tailored to term-term similarity in retrieval contexts rather than document-topic assignment and most of our transcript–subtopic similarity scores fell well below those thresholds, rendering them ineffective for our task. Elekes et al. [21] further demonstrated that general-purpose similarity thresholds do not exist across embedding models, and that thresholds must be determined on a per-model basis.

Given these limitations and in view of the primary focus of this thesis, we turned to prompting large language models (LLMs) for a more robust classification. We tested two state-of-the-art 8-billion-parameter instruction-tuned models, `Qwen3-8B` and `LLaMA-3.1-8B-Instruct`, as zero-shot content classifiers guided through structured few-shot prompting. The prompt, which is provided in Appendix A.3 for reference, framed the task clearly: the model was to classify a transcript snippet of a SQL tutorial video into one or more predefined SQL subcategories, or return an empty list if no subcategory was relevant. The few-shot examples demonstrated three scenarios to the model: videos with one subtopic match, multiple subtopic matches, and no subtopic match at all, to help the model handle these possibilities appropriately. The prompt included not only the task instructions but also the full list of subcategories with associated keywords. The instructions specified that the model could use the presence or semantic meaning of keywords as soft signals and should avoid assigning broad or tangential categories unless they were clearly supported by the transcript content. The model was also required to produce structured output in JSON format enclosed within triple backticks, facilitating direct parsing and analysis. For each video, we provided the video title and either the full transcript or a representative excerpt (truncated at $2,000$ tokens). This LLM-based approach allowed the model to reason over the transcript in light of the provided subtopic compositions and examples, leading to more interpretable and contextually accurate subtopic assignments. The models' outputs, often accompanied by implicit justifications through their alignment with the few-shot guidance, were straightforward to validate programmatically.

Both models outperformed earlier keyword and embedding similarity methods in terms of assignment relevance and consistency. We observed some differences between `Qwen3-8B` and `LLaMA-3.1-8B-Instruct` in terms of precision and granularity of classification, which is detailed in subsection 5.1.4 where an evaluation was conducted to select the better-performing model for final large-scale labeling of the SQL video dataset.

5.1.4. LLM Model Evaluation

Evaluating the accuracy of multi-label topic classification is challenging without an existing ground truth for each video. Because manually annotating thousands of videos across 70 categories was impractical, we used a proxy evaluation method leveraging textbook content as test inputs. Specifically, we randomly sampled thirty index terms from the consolidated textbook keyword list. For each sampled term, we identified a page in one of the source textbooks and extracted a passage discussing that term in context. We then gathered all index terms on that page to identify any additional index terms that accurately reflected the concepts covered in the passage. All verified keywords were mapped to their corresponding subtopics to define the ground truth labels for that sample. This process required manual effort, as although index terms indicate page ranges where a concept is mentioned, they do not specify the exact location or extent of the discussion on the page. As such, automated extraction and validation were not feasible.

We performed the manual validation on 30 randomly sampled cases. Both `Qwen3-8B` and `LLaMA-3.1-8B-Instruct` were tested on these samples, with each model tasked with assigning subtopics to the passage using the structured prompt. An example of one of these evaluation samples is included in Appendix C.1 for reference. As the task allowed multiple valid subtopics per sample, standard multi-label metrics were used to evaluate performance, including micro-averaged precision, recall, and F1-score (see Table 5.2), as well as the confusion matrices summarized in Table 5.3.

Model	Precision	Recall	F1-score	Support
Qwen3-8B	**0.78**	0.73	**0.75**	44
LLaMA-3.1-8B	0.43	**0.75**	0.55	44

Table 5.2: Micro-averaged evaluation results of Qwen3-8B and LLaMA-3.1-8B-Instruct on textbook-derived SQL passages (30 samples, 44 ground truth subtopic labels).

Model	TP	FP	TN	FN
Qwen3-8B	32	**9**	**1,297**	12
LLaMA-3.1-8B	**33**	44	1,262	**11**

Table 5.3: Confusion matrix summary for Qwen3-8B and LLaMA-3.1-8B-Instruct on the evaluation set.

Both models achieved similar recall, as indicated by the confusion matrices and micro-averaged metrics. However, `Qwen3-8B` demonstrated substantially higher precision by producing fewer false positives (9 compared to `LLaMA-3.1-8B`'s 44). This indicates that `Qwen3-8B` was more selective, assigning subtopics only when more confident, whereas `LLaMA-3.1-8B` tended to overpredict, covering broader subtopics at the cost of precision. For example, `LLaMA-3.1-8B` often added tangential subtopics (e.g., predicting additional categories like `Set Operations` or `Logical Connectives` where they were only weakly implied), while `Qwen3-8B`'s outputs generally aligned more closely with the annotated ground truth.

Beyond these quantitative results, we also examined the validity of the predicted subtopics. In this evaluation set, Qwen3-8B achieved a valid prediction rate of 97.56%, with only 2.44% of its more concise predictions corresponding to subtopics not present in the predefined subtopic set. In contrast, LLaMA-3.1-8B had a valid prediction rate of 92.21%, with 7.79% invalid predictions. The invalid subtopics produced by LLaMA-3.1-8B included plausible but non-existent categories such as *Privileges & Roles*, *Assertions in SQL*, *Row-Level Triggers*, *Views*, *Statement-Level Triggers*, and *Trigger Types*. These invalid labels were likely the result of the models inferring labels based on the presence of related keywords provided in the prompt (e.g., *Privileges*, *Triggers*, or *Views*) or those that appeared in the surrounding content or index terms of the sampled texts.

These findings indicate that while both models occasionally assigned subtopics outside the defined set, Qwen3-8B exhibited this behavior less frequently. The higher precision and lower rate of invalid subtopic labels reinforced the decision to select Qwen3-8B as the final classifier for SQL video transcript labeling, suggesting a more reliable and valid subtopic assignment for downstream analysis. To illustrate the model output for real video content, one example of a sound classification result and one example containing representative imperfections are included in Appendix C.2.

5.2. RQ4 Results: Subtopic Coverage Analysis in SQL Videos

The zero-shot Qwen3-8B run produced $9{,}986$ raw subtopic labels for the $4{,}242$ SQL-related videos with transcripts. In total, 300 spurious subtopics that were never defined in the curated list were hallucinated by the model and applied $1,312$ times. These labels were discarded, leaving $8,674$ valid assignments. After filtering, 328 videos (7.73%) ended with an empty label set, typically because the transcripts were either very short (e.g. channel trailers), contained little technical detail (e.g. marketing or motivational content), or focused on tangential material such as installation walk-throughs rather than SQL concepts. Of the $4{,}242$ SQL-related videos that contained transcripts, the remaining $3,914$ classified videos cover 61 of the 70 textbook-derived subtopics, giving an overall coverage of 87.1%. The nine subtopics with zero assignments indicate complete curricular gaps on YouTube within the analyzed sample:

1. *Arity* (arity): No videos explicitly focus on the notion of relation arity (the number of attributes in a relation).
2. *Atomicity & Domains* (atomic domains, atomicity): No videos explicitly focus on atomic domains or the atomicity properties of data values.
3. *Change Tracking & Delta* (delta relation, change relation): No videos explicitly focus on the concepts around delta relations or change relations, used to capture incremental updates.
4. *Difference & EXCEPT* (minus, except all, except clause, except construct, Difference operation, set-difference operation): No videos explicitly focus on SQL's EXCEPT (or MINUS) operator for set-difference.
5. *Example Databases* (banking, university database, sandbox): No videos explicitly focus on canonical example schemas (e.g., banking or university databases) or sandbox environments.
6. *Exceptions & Debugging* (exceptions, exception conditions, sqlstate, debugging, bugs): No videos explicitly focus on handling SQL exceptions, inspecting SQLSTATE codes, or debugging database code.
7. *Operating Systems* (Unix): This concept appears only in the context of certain SQL implementations, most notably PostgreSQL's SIMILAR TO operator, which uses Unix-style regular-expression syntax for more powerful pattern matching than the standard LIKE [80]. No videos explicitly focus on this aspect.
8. *Pointers* (pointers): This concept appears in the context of external procedures and functions written in C, where arguments may be passed as pointers rather than by value to efficiently handle null values, return multiple outputs, and minimize data copying overhead. Such pointer-based interfaces are not supported in “safe” host languages (e.g., Java, C#) that execute within a sandboxed environment, as these languages restrict direct memory manipulation to ensure security and isolation [80]. No videos explicitly cover these low-level pointer mechanics.
9. *Row-Level Security* (Row-level triggers): While triggers were discussed, no content specifically explored row-level security models or trigger-based access control.

The distribution of the 8,674 valid labels is highly skewed. As shown in Table 5.4, the ten most-covered subtopics account for 70.96% of all assignments. These subtopics span core SQL functionality: *Data Manipulation Language (DML)* concerns inserting, deleting, and updating tuples in relations; *Expressions & Syntax* covers SQL's conditional constructs and operators such as `WHERE`, `IN`, and `EXISTS`; *Join Operations* addresses ways of combining relations using various join types and conditions; *Group BY & Having* involves grouping mechanisms and post-group filtering; *Subqueries* focuses on nested and correlated queries; *Aggregate Functions* relates to computing summary statistics over sets of tuples; *Ordering & Limits* concerns result sorting and output restriction; *Data Definition Language (DDL)* covers schema definition and modification; *Select Variants* includes forms and options of the `SELECT` statement; and *Logical Connective* encompasses Boolean operations such as `AND`, `OR`, and `NOT`. Collectively, these subtopics represent the core querying and schema commands that form the foundation of standard SQL usage.

Subtopic	Videos (%)	Median View	Median Like	Median Comment	Median Duration (min)
Data Manipulation Language (DML)	1,335 (31.47%)	5,011	78.5	5	8.75
Expressions & Syntax	882 (20.79%)	4,883	101.5	5	12.12
Join Operations	717 (16.90%)	8,509	188	10	12.02
Group BY & Having	682 (16.08%)	1,986	35.5	3	10.46
Subqueries	650 (15.32%)	5,481	89	6	11.67
Aggregate Functions	582 (13.72%)	666.5	14	1	10.20
Ordering & Limits	410 (9.66%)	3,669.5	90.5	4	12.72
Data Definition Language (DDL)	326 (7.69%)	64,572	1,176	56	35.54
Select Variants	303 (7.14%)	4,828	52	4	7.47
Logical Connectives	268 (6.32%)	9,998	219	12	14.58

Table 5.4: Top 10 SQL subtopics covered in the analyzed YouTube dataset ($N = 4,242$ SQL-related videos with transcripts), showing video count, relative frequency, and median engagement metrics and video duration. Full list is available in Appendix C.3.

While the top 10 subtopics in Table 5.4 dominate coverage, their engagement and duration metrics show considerable variation. For example, *Data Definition Language (DDL)* stands out not only for its high coverage but also for its markedly higher median view count (64,572), likes (1,176), and comments (56), alongside a much longer median duration (35.54 minutes) compared to other frequently covered subtopics such as highest *Data Manipulation Language (DML)* (median 5,011 views, 78.5 likes, 5 comments, 8.75 minutes) or secondary E*xpressions & Syntax* (4,883 views, 101.5 likes, 5 comments, 12.12 minutes). Conversely, several other high-coverage subtopics, such as *Aggregate Functions* and *Group BY & Having*, exhibit lower median view counts (e.g., 666.5 and 1,986, respectively) and shorter durations (10.2 minutes and 10.46 minutes, respectively), suggesting that coverage frequency does not consistently align with higher audience engagement or longer content.

At the long-tail end, as shown in Table 5.5, these are the subtopics that have been assigned the least. Each appears in at most 4 videos, fewer than 0.1% of the sample. *Table Functions* offer a way to treat stored routines as relations, enabling queries to retrieve rows dynamically as if querying a regular table; *Type*, comprising user-defined and structured types, extends SQL's built-in types by allowing the definition of complex, custom data structures; *Set & Assignment* covers operations that assign values to attributes, including the use of the `SET` clause, assignment statements, and setting attributes to `NULL`; *Active Databases* embed triggers and rules to react automatically to changes in data; *Business Logic* facilities capture domain rules directly in stored procedures and constraints; *Fetch/Result APIs* (JDBC, ODBC, ADO.NET, CLI) define how host programs connect to the database and retrieve query results; *Dirty Data* refers to data written by uncommitted transactions, where premature visibility of such data (via dirty reads) can lead to inconsistent or incorrect results; *Common Language Runtime integration (CLR)* in some systems allows developers to write database code, such as stored procedures and functions, using .NET languages; *Sequences* provides a construct to define a sequence by generating unique numeric values; and *Language Integrated Query (LINQ)* brings query syntax directly into host languages like C#, seamlessly embedding SQL-style operations in code. Together, these topics represent the fringes of SQL's capabilities, encompassing advanced extensions, external interfaces, and embedded programming facilities that extend beyond the typically covered core querying and schema commands.

Subtopic	Videos (%)	Median View	Median Like	Median Comment	Median Duration (min)
Table Functions	4 (0.09%)	$22,248.5$	271	3	6.95
Type	4 (0.09%)	23,090	374.5	18	30.20
Set & Assignment	3 (0.07%)	$2,540$	154	3	715.48
Active Databases	3 (0.07%)	$2,540$	154	3	715.48
Business Logic	3 (0.07%)	567	15	0	1.02
Fetch/Result APIs	3 (0.07%)	$41,735$	558	22	21.60
Dirty Data	2 (0.05%)	$32,693$	1,110	69	72.61
Common Language Runtime (CLR)	2 (0.05%)	$1,518$	15.5	3.5	12.07
Sequence	1 (0.02%)	$49,784$	787	31	17.92
Language Integrated Query (LINQ)	1 (0.02%)	$60,855$	161	15	4.57

Table 5.5: Bottom 10 SQL subtopics covered in the analyzed YouTube dataset ($N = 4,242$ SQL-related videos with transcripts), showing video count, relative frequency, and median engagement metrics and video duration. Full list is available in Appendix C.3.

Although the bottom 10 subtopics in Table 5.5 appearing in at most four videos each sometimes display disproportionately high engagement metrics. For example, *Language Integrated Query (LINQ)* $(1$ video) and *Sequence* $(1$ video) both report median view counts above $49,000$, and *Dirty Data* $(2$ videos) shows a median like count exceeding $1,100$. However, these figures are based on extremely small sample sizes and are therefore heavily influenced by individual videos rather than representing broader patterns. Similarly, duration values in the bottom subtopics vary widely, from typical tutorial lengths (e.g., *Table Functions* at 6.95 minutes) to outliers such as *Set & Assignment* and *Active Database* $(715.48$ minutes). These indicate the presence of high audience interest and long content in rare subtopics.

6 Discussion

From the perspective of YouTube dataset construction (RQ1), this study constructed a large-scale dataset of $17,434$ educational YouTube videos related to data systems, using a pipeline grounded in curriculum-aligned search queries and multi-step data cleaning and filtering. The resulting dataset spans a wide range of subtopics within data systems and includes relevant instructional videos and structured information on video structure, context, content, and engagement. Many prior studies focused primarily on specific channels over which the researchers had creator-level access [44, 8]. This approach allowed these studies to obtain deeper viewer profile data or internal analytics not accessible to the general public, but at the cost of dataset breadth. As the scope of their datasets was limited, their focus on a few channels meant they could not reflect the diversity of YouTube educational resources at scale. At the same time, because of the higher-level access, their metadata tended to be more fine-grained and could support deeper analysis on specific audiences or creator-side metrics. Other studies concentrated on features of videos that are difficult to collect automatically, such as multimodal presentation characteristics or cognitive signals [9, 93, 89]. Those studies typically relied on small-scale, manually collected and annotated samples, due to the labor-intensive nature of such data gathering. In cases where certain metadata was not directly accessible to the public, some studies supplemented their datasets by obtaining proxy data from commercial marketing analytics services, such as sharing metrics, to approximate user interaction patterns [89].

Some large-scale collection and analysis of YouTube educational data in STEM fields are at a general science level, as reflected in works such as Shaikh et al. [77] and Debove et al. [15]. Shaikh et al. focused on videos that cite research articles, aiming to explore their societal and scholarly impact. Their dataset was broad across scientific domains but limited to videos mentioning scholarly outputs, without aiming at any specific field. Debove et al. conducted a survey and metadata analysis of French science communication channels, examining their characteristics, institutional background, and communicator goals. However, their work did not target any specific academic subfield, nor did it attempt curriculum-aligned data collection.

In the computer science domain, some work [36] primarily focused on analyzing popularity trends in programming tutorials using video-level metadata (e.g., views, upload dates) across multiple programming languages. Some work [5] applied machine learning techniques to classify videos into broad computer science subfields (e.g., computer hardware, computer networks) based on subtitle keyword extraction. However, their effort remained at a task-specific dataset and did not aim for curricular coverage or dataset reusability.

In this context, the dataset constructed in this study aimed to collect all publicly accessible YouTube educational resources within a well-defined and specialized field, data systems in computer science, at a level of granularity aligned with curriculum topics, and ensure broad topical coverage within this area. Efforts were made to preserve as much potentially useful metadata as possible so that the dataset could serve not only the immediate objectives of this research but also support future investigations with varied interests. The scale, transparency, and reproducibility of the data pipeline are intended to complement prior work by providing a resource that enables repeatable analyses and facilitates further exploration of educational YouTube content in this domain.

Regarding the key characteristics of the dataset (RQ2), this study provides a descriptive overview of educational YouTube videos related to data systems, including their volume, engagement patterns, transcript availability, language distribution, and geographic origins. All key interaction metrics, view counts, like counts, comment counts, channel total views, and channel subscriber counts, exhibit power-law distributions, where a small fraction of videos and channels attract the vast majority of attention. This dynamic aligns with known patterns on social platforms [61]. The pronounced gap between view counts and comment/like counts suggests that data systems educational YouTube videos tend to function more as passive content consumption resources rather than dialogic or interactive spaces. Data systems videos and their hosting channels experienced marked growth in 2020, coinciding with the educational shift to online formats triggered by the COVID-19 pandemic. An earlier increase in channel creation around 2011 may relate to the rise of smartphones, mobile internet access, and the early wave of MOOCs. In recent years, while the annual number of new videos has remained at a high level with some fluctuations, the number of new contributing channels has declined year by year, possibly reflecting the consolidation of content creation into established producers and higher barriers to entry for new creators.

Only about 6% of videos had creator-provided or reviewed captions, while around 80% relied on auto-generated captions, and the remaining 14% lacked usable transcripts due to poor audio quality or creator-disabled captioning. This reflects a general underinvestment by creators in accessibility features, despite evidence that creator-uploaded captions are linked to stronger audience engagement. Moreover, language attributes were often missing or incomplete in the metadata, limiting a full understanding of language trends. Nevertheless, some underrepresented languages, either in audio or text, showed high audience engagement, hinting at unmet local demand in data systems education. The geographic distribution reinforces this point with India and the United States as the main sources of data systems educational content, but countries like Brazil and the United Kingdom contribute content with comparatively more active communities. Regional variation in content volume and engagement suggests differences in production incentives, platform strategies, or audience demand.

The distribution of data systems topics shows similar trends. Most creations focused on SQL, database normalization, and data mining. This concentration likely reflects a combination of user demand, the accessibility of these topics to beginner audiences, and the amplification effect of YouTube's recommendation algorithm. Meanwhile, topics like data visualization, data security and access management, distributed database management systems, and object-oriented data models, though less represented, achieved higher engagement, indicating possible areas of learner interest that exceed the available supply. Some advanced but less frequent topics, such as Object-Oriented Data Models, Relational Theory, and Semi-Structured Data Models, are highly theoretical, conceptually abstract, and typically consist of stable foundational knowledge that may not readily stimulate the production of new content.

From the audience engagement modeling (RQ3), several key patterns emerge regarding the factors that shape how data systems educational videos on YouTube capture viewer interest. First, longer videos were associated with higher engagement, suggesting that despite pedagogical recommendations favoring instructional units shorter than 6 minutes [28, 11], viewers in this domain may seek more comprehensive explanations that require longer runtime. This finding aligns with the descriptive statistics of the dataset: the videos had a median duration of 10.03 minutes, with an interquartile range of 4.82 to 21.96 minutes, indicating videos are generally longer than 6 minutes. These suggest that creators should balance brevity with the depth required to adequately cover technical topics like data systems, rather than adhering rigidly to generalized guidelines on video length.

Channel-level indicators, subscriber count, and total view count were positive predictors of engagement. This aligns with findings from Velho et al. [95] and Bello-Bravo et al. [8], both of which noted the "rich-get-richer" dynamic where established channels accrue more engagement, partly through platform recommendation algorithms, which implies that newer or smaller creators face structural disadvantages. However, higher channel productivity, defined as the frequency of uploads relative to channel age, was negatively associated with engagement, as was channel age itself. These patterns imply that simply producing more content or being active on the platform for longer does not guarantee sustained or increasing engagement; creators should focus on producing content that meets learner needs and platform dynamics rather than sheer volume.

Moreover, the topic of the video itself influenced engagement. The presence of SQL-related content was strongly associated with higher audience interaction, confirming SQL's centrality within data systems education and its continued appeal to learners, which suggests that focusing on high-demand, foundational topics can be an effective engagement strategy.

Additionally, title length had a small but negative association with engagement, indicating that longer or more complex titles might slightly deter viewer interaction. Clear, concise titling may help improve click-through and engagement rates, aligning with best practices in educational media communication.

Language and regional attributes revealed more nuanced patterns. While most language and country indicators did not significantly predict engagement, videos labeled with the textual language "English (India)" showed notably higher engagement. While the linguistic content may not differ markedly from standard English, this label likely functions as a proxy for regionally localized content tailored to Indian audiences, such as videos created by Indian educators, using local examples or communication styles familiar to learners in the region. Given India's large and highly active user base on YouTube, such videos may benefit from both cultural resonance and algorithmic amplification, leading to enhanced engagement. On the cultural side, a study shows that YouTube actively aligns itself with regional language markets in India by foregrounding content that resonates with linguistic and cultural identities [59]. On the algorithmic side, Covington et al. [13] detailed how YouTube's recommendation engine uses deep neural networks trained on user behavior signals, including watch history, geolocation, and language preferences, to personalize recommendations at scale. Videos that match a user's regional and linguistic profile are more likely to be recommended and promoted through home feeds and autoplay, increasing their exposure and engagement potential. This points to an opportunity for targeted content development or localization efforts aimed at audiences of different cultural backgrounds, and it hints at regional preferences that deserve further exploration. The lack of significant effects for other language and country variables may reflect the limitations of available metadata or the need to consider deeper contextual factors beyond self-declared tags.

Other factors, such as valence, density, readability, and most language and country attributes, did not show any significant association with engagement. This contrasts with Fischer et al. [24], who found affective characteristics significant in driving engagement for TED Talks. The divergence may reflect differences in content type: while affective tone matters in science communication aimed at broad public audiences like TED Talks, it appears less influential in specialized technical education, where structural and informational qualities may dominate.

From the analysis of subtopic coverage in SQL-related YouTube videos (RQ4), several key findings emerge that highlight both the strengths and gaps of current educational content. First, the overall subtopic coverage is reasonably broad: 87.1% of the textbook-derived SQL subtopics are represented across the sampled videos, indicating that YouTube offers substantial material spanning core aspects of SQL. However, the distribution of subtopic coverage is highly skewed, with the top 10 subtopics, such as *Data Manipulation Language (DML)*, *Expressions & Syntax*, *Join Operations*, and *Subqueries*, accounting for over 70% of all valid label assignments. This dominance of core querying and schema commands suggests that YouTube content largely aligns with beginner and intermediate learning priorities, focusing on practical tasks that directly support common SQL use cases.

Conversely, certain subtopics received little or no attention at all. Notably, subtopics, such as *Arity*, *Atomicity & Domains*, *Change Tracking & Delta*, *Difference & EXCEPT*, *Exceptions & Debugging*, *pointers*, and *Row-Level Security*, related concepts, were entirely absent in the labeled dataset. These omissions reveal content gaps, particularly in areas of foundational data model concepts, advanced database management features, security mechanisms, and systems-level implementation details that are less frequently addressed in instructional materials. While YouTube may serve as a rich source for essential SQL instruction, this is not the case for more advanced or theoretical topics. Learners who want to go deeper into the language may find it difficult to locate much accessible content on these areas. This also signals an opportunity for educators, institutions, and creators to target these neglected areas, thereby enriching the ecosystem of openly accessible educational resources.

Similarly, the rare subtopics, spanning integration, extension, and procedural mechanisms, remain notably underrepresented in SQL-related YouTube content. These include, among others, *Table Func-*

tions, *Type*, *Set & Assignment*, *Active Databases*, *Business Logic*, *Fetch/Result APIs*, *Dirty Data*, *Common Language Runtime (CLR)*, *Sequence*, and *Language Integrated Query (LINQ)*. While these features are technically significant as they support advanced modeling, automation, and system-level integration, they are often tied to specific platforms or require interaction with external programming environments. Their niche applicability, steeper learning curves, and lower demand from entry-level audiences might make them less compatible with the general-purpose tutorial format that dominates the platform. Much of the SQL-related content on YouTube appears to be created toward practical, industry-oriented learning goals, such as interview preparation, job-focused upskilling, or hands-on demonstrations with specific DBMS tools. In such contexts, creators may prioritize widely used SQL constructs that align with hiring expectations or certification standards, rather than exploring more theoretical or specialized extensions. This emphasis on practicality contributes to the skewed representation of SQL knowledge, where portable, query-centric skills dominate, while leaving gaps for more advanced or academic aspects of the language.

7 Limitations

While the dataset offers broad coverage of data systems educational content on YouTube, its design is shaped by the use of publicly accessible metadata from the YouTube Data API. This approach enables reproducibility and transparency but excludes proprietary metrics such as detailed audience demographics or precise watch-time analytics. The dataset is thus more suited for analyses focused on structural, linguistic, and content-level characteristics, rather than fine-grained viewer behavior or creator-side performance indicators typically available only through privileged creator access. Additionally, while search queries were carefully crafted to ensure wide topical coverage, their reach was inherently influenced by both query phrasing and YouTube's search-matching behavior. Some relevant videos may not have been retrieved, such as those with unconventional titles or tags, or if the algorithm failed to surface content due to misalignment with the query terms.

The dataset captures a snapshot of publicly available content as of December 24, 2024. Given the dynamic nature of YouTube, where videos may be deleted, made private, or affected by policy enforcement, some content may become unavailable over time. While this static snapshot enables consistent and replicable analysis, it reflects the platform's state at a specific point in time and may not fully represent longer-term trends or content stability.

The relevance filtering process employed a 7B-parameter embedding model, chosen as a balance between performance and resource constraints. While effective for large-scale classification, this setup may not fully explore the marginal benefits of more advanced or larger-scale models, such as the potential gains in embedding quality and classification accuracy.

In addition to dataset-related limitations, some constraints specific to the regression analysis should be noted. First, although the model explained approximately 35% of the variance in engagement, this leaves a substantial proportion of variance unaccounted for. This is consistent with the inherent complexity of audience behavior on social platforms and the limited set of features used in this study. The model's covariates included structural, linguistic, and channel-level metrics that were publicly accessible through the YouTube API and transcript processing. More nuanced features requiring manual annotation, such as visual style, cognitive signals, or instructional style, were not in the scope of this study. This limits the interpretability of the model regarding presentation factors that could affect educational impact. In addition, the lack of declarations in some videos and the potential mismatch between declared and actual language or location may have weakened the model's ability to detect associations involving linguistic or geographic context.

Some predictor variables, such as subscriber count and total channel views, exhibited multicollinearity, which posed challenges in isolating their unique contributions. While variance inflation factors were monitored and coefficients remained interpretable, the overlap between these metrics reflects a common difficulty in distinguishing between closely correlated indicators of creator popularity and reach. The joint presence of subscriber count and total channel views in the model may result in shared explanatory power, making it difficult to determine which of the two is the stronger independent driver of engagement. The findings imply that those two channel metrics should be interpreted as indicating a general effect of creator popularity, rather than distinct effects of each metric.

Finally, the analysis was inherently correlational. As with all observational studies on platform data, causal inferences about the effects of video features on engagement cannot be drawn from these results alone. Unobserved confounders, such as external promotion or recommendation algorithm dynamics, likely influenced outcomes.

Apart from the dataset- and model-level limitations already discussed, the subtopic coverage analysis in RQ4 involved some methodological trade-offs. Subtopic labels were assigned using a zero-shot large language model applied to video transcripts, enabling efficient large-scale classification without manual annotation. However, this approach could be inherently sensitive to ambiguity in terminology and conceptual overlap between subtopics. Although spurious labels were filtered and assignments reviewed at the aggregate level, individual label assignments were too many to be manually verified per assignment, which may introduce susceptibility to hallucinations, omission, or misclassification for subtopics. Apart from that, because the prompting strategy relied on a mapping of index keywords to subtopics, content with few or vague lexical cues may have been harder for the model to identify reliably, as well as subtopics lacking strong or distinctive keywords might therefore be more prone to being overlooked or incorrectly assigned.

Another limitation is that the SQL subtopic coverage analysis was necessarily limited to videos with transcripts, and the quality of those transcripts may have influenced subtopic detection. Videos with incomplete, inaccurate, or low-quality transcripts (for example, due to poor auto-captioning or unclear speech) were less likely to have subtopics assigned or were labeled less reliably. The results, therefore, may underrepresent videos where relevant SQL concepts were present but not adequately captured in text form. Similarly, while the subtopic list was derived from standard curriculum sources, it may not fully reflect the varied ways in which SQL concepts are described or taught informally by creators on YouTube. Thus, content using alternative terminology or non-standard framing could be underrepresented in the analysis.

8 Ethical Considerations

This chapter outlines the ethical considerations undertaken in the design and execution of this study, including how data was collected, processed, and analyzed responsibly. Although the research focuses on publicly available content, particular care was taken to ensure that all steps adhered to principles of transparency, fairness, and respect for the individuals and communities represented in the dataset.

8.1. Data Management

The dataset used in this study was constructed entirely from publicly accessible metadata and transcripts of YouTube videos related to data systems education. All data was obtained via official APIs provided by YouTube, including the retrieval of video transcripts using the `youtube-transcript-api` library, which internally accesses YouTube's official transcript API endpoints.

Importantly, the dataset contains no personal identifiers or sensitive information related to user privacy. All collected attributes were already openly available to all users of YouTube at the time of data retrieval. The dataset explicitly excludes any personal information about individual viewers or commenters, and no attempts were made to collect, infer, or analyze personal or demographic data from users.

The analysis conducted within this research exclusively focuses on aggregated trends and educational characteristics pertinent to the data systems domain. At no point were individual users or channels tracked, profiled, or linked to external datasets. The sole analytical aim was to examine content-level patterns, engagement metrics, and topical coverage relevant to educational use in data systems.

To clarify the intent and scope of use, a clear academic use statement was included alongside the open dataset[1]. This statement specifies that the data was collected exclusively for non-commercial, research purposes and is composed only of content already publicly accessible on YouTube. This is to ensure transparency and to explicitly reject any intention to misuse, redistribute, or facilitate misuse of platform content in ways that would violate intellectual property rights or individual privacy.

8.2. Responsible Use of Automated Methods

This study makes use of automated techniques, including large language models and embedding-based classifiers, in tasks such as video relevance filtering and subtopic classification. While these methods enable the reduction of manual labeling effort and large-scale analysis, particular care was taken to apply them responsibly.

In the video relevance classification pipeline, synthetic training labels were generated using multiple prompt strategies with an LLM, which were evaluated on a manually annotated subset of videos to balance high recall of relevant content with a minimized false positive rate. Following label generation, a binary classifier was trained under class imbalance conditions. SMOTE was applied to the training data to mitigate class imbalance and improve the model's ability to recognize underrepresented irrelevant

[1] `https://doi.org/10.17605/OSF.IO/FTN2S`

cases. Multiple performance metrics were reported for both classes to assess class-specific performance. Cost-sensitive thresholding was also applied to better control for the asymmetric cost of false positives and false negatives.

For SQL subtopic classification, LLMs were used to group dispersed textbook-derived index terms into coherent subtopics. This grouping process was not fully automated: the initial groupings generated by the LLM were manually reviewed and adjusted to ensure pedagogical consistency and topical clarity. To assess the reliability of using LLMs to assign videos to fine-grained subtopics, especially in the absence of large-scale, human-labeled ground truth data, an evaluation was conducted in which textbook passages corresponding to specific subtopics were compiled and used as inputs to the same classification pipeline. The model's ability to correctly classify these known instructional texts served as an external validation of its behavior.

These design choices reflect a broader commitment to ethical use of automated tools in research. By validating model behavior against human-labeled samples, applying corrective techniques such as SMOTE, and transparently reporting class-specific performance, the study minimizes the risk of unintended bias, misclassification, or misuse. Particular attention was given to class imbalance, false positive risk, and domain-specific consistency to reduce the amplification of systemic or model-induced biases. Manual oversight in key stages, such as prompt selection, subtopic grouping, and evaluation without ground truth, ensures that automated methods remain aligned with the research's educational focus and do not compromise interpretability, fairness, or integrity.

8.3. Reflection on Research Impact and Fairness

This study analyzes patterns of engagement and topical coverage in YouTube's educational videos on data systems. While no direct interaction with human subjects occurred, the research proposes implications for creators, platforms, and learners. For example, associations identified between engagement and video features may inform content strategies or affect how material is surfaced by recommendation systems. To avoid overgeneralization or misinterpretation, several steps were taken during analysis. Descriptive statistics and model results were reported alongside variance, outlier effects, and class imbalance considerations. Engagement trends were interpreted in relative rather than absolute terms, and no normative claims were made about what content "should" be promoted. Language, geography, and topic distributions were analyzed with attention to platform dynamics and content availability, not as indicators of creator or learner intent.

This study also highlights disparities in topic representation and language coverage within educational content on YouTube, such as the overrepresentation of beginner-level SQL content and the dominance of English-language material, which may shape inequalities in informal educational resource access. The work aims to document these patterns and contribute empirical evidence to inform future efforts toward improving content diversity, accessibility, and topical balance on open learning platforms.

8.4. Transparency and Reproducibility

This study was designed with an emphasis on transparency and reproducibility. All core methods, such as the data collection pipeline, filtering processes, prompt design, model configurations, and classification strategies, are described in detail within the thesis. Evaluation metrics and key results are also included to support interpretability and traceability. Furthermore, the related code and dataset have been published openly on the Open Science Framework (OSF)[2], ensuring that other researchers can independently verify, replicate, and build upon this work.

[2] https://doi.org/10.17605/OSF.IO/FTN2S

9 Conclusion

This thesis presents a comprehensive investigation into educational content on YouTube within the domain of data systems. Motivated by the growing influence of informal online learning platforms and the relative scarcity of structured analyses in specialized technical fields on YouTube, the study addresses four core research questions regarding the availability, characteristics, engagement dynamics, and topical distribution of educational videos. A curated dataset of 17,434 instructional YouTube videos on data systems was constructed using curriculum-informed search queries and multilevel metadata collection. A filtering pipeline leveraging large language models and embedding-based classification was employed to retain relevant educational videos. Using this dataset, descriptive analyses mapped the landscape of available material and revealed disparities across content volume, engagement levels, transcript availability, language use, and geographical origin. Statistical engagement modeling identified potential features associated with audience engagement, based on a diverse set of structural, linguistic, and contextual attributes derived from video and channel-level metadata. By classifying video content against textbook-based subtopics, using SQL as a case study, the analysis identified patterns of both overrepresented and underrepresented SQL themes. Overall, this work charted the types and characteristics of data systems educational videos on YouTube, examined factors associated with their engagement, and evaluated the coverage of YouTube content under a fine-grained, textbook-aligned topic structure.

10

Future Work

A potential next step is to improve and expand the dataset of educational videos, both in breadth and depth. One opportunity is to broaden the dataset's language coverage. Although our current collection includes videos in multiple languages, the search queries used were exclusively in English, and it remains unclear how effective YouTube's search is at retrieving multilingual content based on English queries. Future work could therefore incorporate multilingual search queries and localized data collection strategies to capture educational videos that may have been missed due to language limitations. This expansion would enable a more robust analysis of cross-linguistic differences in educational content and user engagement. Notably, our data suggest that some under-represented languages may host disproportionately engaged learning communities. Such findings point to rich educational ecosystems that benefit deeper investigation.

Moreover, real-time and continual data collection can be considered to keep the dataset up-to-date. The landscape of YouTube content is highly dynamic, and millions of new videos are uploaded daily. A static snapshot collected at one time will soon become outdated as new educational videos, channels, and trends emerge. Future work can implement an automated pipeline to periodically query the YouTube API and refresh the dataset with new videos and updated engagement metrics. Such a live dataset would enable a longitudinal analysis of how educational content evolves. It also opens the possibility for real-time monitoring of audience responses to current educational videos. Additionally, improving data quality through richer metadata is worthwhile, for example, retrieving higher-resolution transcripts transcribed by state-of-the-art transcription models for videos with auto-generated captions or those initially missing transcripts, and capturing multimodal features like embodied modes and filmic modes [93, 9]. Ensuring the dataset remains comprehensive, multilingual, current, and multimodal will provide a stronger foundation for all subsequent analyses.

While this thesis focused on measuring engagement at the video level, future work could incorporate learner-centric outcomes. This could involve surveys, interviews, or behavioral studies to assess how learners perceive educational quality, usefulness, or learning gains from YouTube videos, as previous studies have practiced [22]. Integrating feedback loops from actual learners would enrich the understanding of educational effectiveness beyond engagement metrics. Moreover, analyzing the sentiment and argumentative expression of viewer comments could reveal how positively or negatively learners react to a video, and measure audience attitudes and satisfaction, complementing quantitative metrics beyond the count of views, likes, and comments [17, 46].

The current SQL subtopic coverage study revealed both covered and neglected areas. Future research could extend this work by analyzing concept-conveying strategies in different subtopics. Previous studies have revealed the difficulty of different SQL constructs [3] and identified the types, causes, and persistence of student errors [4, 86, 100]. Some have also explored knowledge transfer difficulties across different database query languages [48]. Building on this, our dataset of classified SQL instructional videos provides a source to investigate how knowledge transfer is dealt with and how complex concepts are explained among the creators. Notional machines like visualizations and analogies could be applied to help learners build mental models for understanding abstract concepts [23]. Researchers have studied how the concept of variables is introduced and explained in online courses of programming education, often using metaphors like ”variables as boxes” to aid conceptual understanding [94].

While this study focuses on data systems, many other computer science domains, such as programming languages, algorithms, data structures, and web development, also have rich educational ecosystems on YouTube. Future research could replicate and adapt the methodology to map and analyze content coverage and engagement patterns across these domains. A comparative analysis of fields might reveal differences in learner preferences, content formats, or creation strategies. Additionally, broadening the scope enables cross-domain insights, as we might discover best practices in one field that could benefit content creation in another.

References

[1] Acm Computing Curricula Task Force, ed. *Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science*. en. ACM, Inc, Jan. 2013. ISBN: 978-1-4503-2309-3. DOI: 10.1145/2534860. URL: http://dl.acm.org/citation.cfm?id=2534860 (visited on 07/05/2025).

[2] Willem Aerts, George Fletcher, and Daphne Miedema. "A Feasibility Study on Automated SQL Exercise Generation with ChatGPT-3.5". In: *Proceedings of the 3rd International Workshop on Data Systems Education: Bridging education practice with education research*. DataEd '24. New York, NY, USA: Association for Computing Machinery, July 2024, pp. 13–19. ISBN: 979-8-4007-0678-3. DOI: 10.1145/3663649.3664368. URL: https://dl.acm.org/doi/10.1145/3663649.3664368 (visited on 06/23/2025).

[3] Alireza Ahadi, Julia Prior, Vahid Behbood, and Raymond Lister. "A Quantitative Study of the Relative Difficulty for Novices of Writing Seven Different Types of SQL Queries". In: *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*. ITiCSE '15. New York, NY, USA: Association for Computing Machinery, June 2015, pp. 201–206. ISBN: 978-1-4503-3440-2. DOI: 10.1145/2729094.2742620. URL: https://doi.org/10.1145/2729094.2742620 (visited on 06/23/2025).

[4] Alireza Ahadi, Julia Prior, Vahid Behbood, and Raymond Lister. "Students' Semantic Mistakes in Writing Seven Different Types of SQL Queries". In: *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*. ITiCSE '16. New York, NY, USA: Association for Computing Machinery, July 2016, pp. 272–277. ISBN: 978-1-4503-4231-5. DOI: 10.1145/2899415.2899464. URL: https://doi.org/10.1145/2899415.2899464 (visited on 06/23/2025).

[5] Pooja Ajwani and Harshal A. Arolkar. "Classification of Domains in Computer Science Using Random Forest Algorithm for YouTube Dataset". en. In: *Second International Conference on Image Processing and Capsule Networks*. ISSN: 2367-3389. Springer, Cham, 2022, pp. 662–670. ISBN: 978-3-030-84760-9. DOI: 10.1007/978-3-030-84760-9_56. URL: https://link.springer.com/chapter/10.1007/978-3-030-84760-9_56 (visited on 07/12/2025).

[6] Jatin Ambasana, Sameer Sahasrabudhe, and Sridhar Iyer. "SQL-Wordle: Gamification of SQL Programming Exercises". In: *Proceedings of the ACM Conference on Global Computing Education Vol 2*. CompEd 2023. New York, NY, USA: Association for Computing Machinery, Dec. 2023, p. 190. ISBN: 979-8-4007-0374-4. DOI: 10.1145/3617650.3624949. URL: https://dl.acm.org/doi/10.1145/3617650.3624949 (visited on 06/24/2025).

[7] Solomon R. Antony and Dinesh Batra. "CODASYS: a consulting tool for novice database designers". In: *SIGMIS Database* 33.3 (Aug. 2002), pp. 54–68. ISSN: 0095-0033. DOI: 10.1145/569905.569911. URL: https://doi.org/10.1145/569905.569911 (visited on 06/23/2025).

[8] Julia Bello-Bravo, Jane Payumo, and Barry Pittendrigh. "Measuring the impact and reach of informal educational videos on YouTube: The case of Scientific Animations Without Borders". English. In: *Heliyon* 7.12 (Dec. 2021). Publisher: Elsevier. ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2021.e08508. URL: https://www.cell.com/heliyon/abstract/S2405-8440(21)02611-6 (visited on 09/27/2024).

[9] Edgar Bernad-Mechó and Julia Valeiras-Jurado. "Multimodal engagement strategies in science dissemination: A case study of TED talks and YouTube science videos". en. In: *Discourse Studies* 25.6 (Dec. 2023). Publisher: SAGE Publications, pp. 733–754. ISSN: 1461-4456. DOI: 10.1177/14614456231161755. URL: https://doi.org/10.1177/14614456231161755 (visited on 09/27/2024).

[10] Douglas B. Bock and Susan E. Yager. “Improving Entity Relationship Modeling Accuracy with Novice Data Modelers”. In: *Journal of Computer Information Systems* 42.2 (Jan. 2002). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08874417.2002.11647489, pp. 69–75. ISSN: 0887-4417. DOI: 10.1080/08874417.2002.11647489. URL: https://doi.org/10.1080/08874417.2002.11647489 (visited on 06/23/2025).

[11] Cynthia J. Brame. “Effective Educational Videos: Principles and Guidelines for Maximizing Student Learning from Video Content”. In: *CBE—Life Sciences Education* 15.4 (Dec. 2016). Publisher: American Society for Cell Biology (lse), es6. DOI: 10.1187/cbe.16-03-0125. URL: https://www.lifescied.org/doi/10.1187/cbe.16-03-0125 (visited on 06/24/2025).

[12] Thomas M Connolly and Carolyn E Begg. “A Constructivist-Based Approach to Teaching Database Analysis and Design”. en. In: *Journal of Information Systems Education* 17.1 (2006), pp. 43–54.

[13] Paul Covington, Jay Adams, and Emre Sargin. “Deep Neural Networks for YouTube Recommendations”. In: *Proceedings of the 10th ACM Conference on Recommender Systems*. RecSys ’16. New York, NY, USA: Association for Computing Machinery, Sept. 2016, pp. 191–198. ISBN: 978-1-4503-4035-9. DOI: 10.1145/2959100.2959190. URL: https://dl.acm.org/doi/10.1145/2959100.2959190 (visited on 07/12/2025).

[14] N.H.T.M. De Siva and R.A.H.M. Rupasingha. “Classifying YouTube Videos Based on Their Quality: A Comparative Study of Seven Machine Learning Algorithms”. In: *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*. ISSN: 2164-7011. Aug. 2023, pp. 251–256. DOI: 10.1109/ICIIS58898.2023.10253580. URL: https://ieeexplore.ieee.org/document/10253580/references (visited on 07/12/2025).

[15] Stéphane Debove, Tobias Füchslin, Tania Louis, and Pierre Masselot. “French Science Communication on YouTube: A Survey of Individual and Institutional Communicators and Their Channel Characteristics”. English. In: *Frontiers in Communication* 6 (Apr. 2021). Publisher: Frontiers. ISSN: 2297-900X. DOI: 10.3389/fcomm.2021.612667. URL: https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2021.612667/full (visited on 09/27/2024).

[16] César Domínguez and Arturo Jaime. “Database design learning: A project-based approach organized through a course management system”. In: *Comput. Educ.* 55.3 (Nov. 2010), pp. 1312–1320. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2010.06.001. URL: https://doi.org/10.1016/j.compedu.2010.06.001 (visited on 06/24/2025).

[17] Ilana Dubovi and Iris Tabak. “Interactions between emotional and cognitive engagement with science on YouTube”. en. In: *Public Understanding of Science* 30.6 (Aug. 2021). Publisher: SAGE Publications Ltd, pp. 759–776. ISSN: 0963-6625. DOI: 10.1177/0963662521990848. URL: https://doi.org/10.1177/0963662521990848 (visited on 09/28/2024).

[18] Kavisha Duggal, Anukool Srivastav, and Satvinder Kaur. “Gamified Approach to Database Normalization”. en. In: *International Journal of Computer Applications* 93.4 (May 2014). Publisher: Foundation of Computer Science (FCS), pp. 47–53. ISSN: 0975-8887. DOI: 10.5120/16207-5505. URL: https://scispace.com/papers/gamified-approach-to-database-normalization-4g90q1mj1e (visited on 06/24/2025).

[19] Lucila Dughera, Fernando Bordignon, and Esteban Azzara. “A literature review of the YouTube phenomenon and the teaching and learning practices”. In: Feb. 2021.

[20] Ian Duncan, Lee Yarwood-Ross, and Carol Haigh. “YouTube as a source of clinical skills education”. In: *Nurse Education Today* 33.12 (Dec. 2013), pp. 1576–1580. ISSN: 0260-6917. DOI: 10.1016/j.nedt.2012.12.013. URL: https://www.sciencedirect.com/science/article/pii/S0260691712004108 (visited on 07/12/2025).

[21] Ábel Elekes, Adrian Englhardt, Martin Schäler, and Klemens Böhm. “Toward meaningful notions of similarity in NLP embedding models”. en. In: *International Journal on Digital Libraries* 21.2 (June 2020), pp. 109–128. ISSN: 1432-1300. DOI: 10.1007/s00799-018-0237-y. URL: https://doi.org/10.1007/s00799-018-0237-y (visited on 05/21/2025).

[22] Danelly Susana Esparza Puga and Mario Sánchez Aguilar. "Students' Perspectives on Using Youtube as a Source of Mathematical Help: The Case of 'Julioprofe'". en. In: *International Journal of Mathematical Education in Science and Technology* 54.6 (2023). Publisher: Taylor & Francis ERIC Number: EJ1387286, pp. 1054–1066. ISSN: 0020-739X. DOI: 10.1080/0020739X.2021.1988165. (Visited on 09/24/2024).

[23] Sally Fincher, Johan Jeuring, Craig S. Miller, Peter Donaldson, Benedict du Boulay, Matthias Hauswirth, Arto Hellas, Felienne Hermans, Colleen Lewis, Andreas Mühling, Janice L. Pearce, and Andrew Petersen. "Notional Machines in Computing Education: The Education of Attention". In: *Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education*. ITiCSE-WGR '20. New York, NY, USA: Association for Computing Machinery, Dec. 2020, pp. 21–50. ISBN: 978-1-4503-8293-9. DOI: 10.1145/3437800.3439202. URL: https://dl.acm.org/doi/10.1145/3437800.3439202 (visited on 07/08/2025).

[24] Olivia Fischer, Loris T. Jeitziner, and Dirk U. Wulff. "Affect in science communication: a data-driven analysis of TED Talks on YouTube". en. In: *Humanities and Social Sciences Communications* 11.1 (Jan. 2024). Publisher: Palgrave, pp. 1–9. ISSN: 2662-9992. DOI: 10.1057/s41599-023-02247-z. URL: https://www.nature.com/articles/s41599-023-02247-z (visited on 09/22/2024).

[25] V. Gigoryeva-Golubeva, E. Silina, and E. Surinova. "YouTube English video lectures as a basis of CLIL classes for students of mathematics". en. In: *Journal of Physics: Conference Series* 1691.1 (Nov. 2020). Publisher: IOP Publishing, p. 012045. ISSN: 1742-6596. DOI: 10.1088/1742-6596/1691/1/012045. URL: https://dx.doi.org/10.1088/1742-6596/1691/1/012045 (visited on 09/27/2024).

[26] Maarten Grootendorst. *KeyBERT: Minimal keyword extraction with BERT.* Version v0.3.0. 2020. DOI: 10.5281/zenodo.4461265. URL: https://doi.org/10.5281/zenodo.4461265.

[27] Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. "FullStop: Multilingual Deep Models for Punctuation Prediction". In: (June 2021). URL: http://ceur-ws.org/Vol-2957/sepp_paper4.pdf.

[28] Philip J. Guo, Juho Kim, and Rob Rubin. "How video production affects student engagement: an empirical study of MOOC videos". In: *Proceedings of the first ACM conference on Learning @ scale conference*. L@S '14. New York, NY, USA: Association for Computing Machinery, Mar. 2014, pp. 41–50. ISBN: 978-1-4503-2669-8. DOI: 10.1145/2556325.2566239. URL: https://dl.acm.org/doi/10.1145/2556325.2566239 (visited on 06/24/2025).

[29] Haryanto, Wahyu Mustafa Kusuma, Farid Mutohhari, Muhammad Nurtanto, and Suyitno Suyitno. "Innovation Media Learning: Online Project-Based Learning (O-PBL) on Drawing Competence in Automotive Engineering Using Video on YouTube". en. In: *Journal of Physics: Conference Series* 2111.1 (Nov. 2021). Publisher: IOP Publishing, p. 012020. ISSN: 1742-6596. DOI: 10.1088/1742-6596/2111/1/012020. URL: https://dx.doi.org/10.1088/1742-6596/2111/1/012020 (visited on 09/22/2024).

[30] Vanessa M. Hill, Will J. Grant, Melanie L. McMahon, and Isha Singhal. "How prominent science communicators on YouTube understand the impact of their work". English. In: *Frontiers in Communication* 7 (Dec. 2022). Publisher: Frontiers. ISSN: 2297-900X. DOI: 10.3389/fcomm.2022.1014477. URL: https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2022.1014477/full (visited on 09/27/2024).

[31] Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, Haofen Wang, Jun Yu, and Min Zhang. *KaLM-Embedding: Superior Training Data Brings A Stronger Embedding Model*. arXiv:2501.01028 [cs]. Jan. 2025. DOI: 10.48550/arXiv.2501.01028. URL: http://arxiv.org/abs/2501.01028 (visited on 03/14/2025).

[32] Tauqeer Hussain. "Teaching Entity-Relationship Models Effectively". In: *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*. Dec. 2016, pp. 264–269. DOI: 10.1109/CSCI.2016.0058. URL: https://ieeexplore.ieee.org/document/7881351 (visited on 06/23/2025).

[33] Yi-Ling Hwong, Carol Oliver, Martin Van Kranendonk, Claude Sammut, and Yanir Seroussi. “What makes you tick? The psychology of social media engagement in space science communication”. en. In: *Computers in Human Behavior* 68 (Mar. 2017), pp. 480–492. ISSN: 07475632. DOI: 10.1016/j.chb.2016.11.068. URL: https://linkinghub.elsevier.com/retrieve/pii/S0747563216308172 (visited on 06/26/2025).

[34] Mustafa I. Eid. “A Learning System For Entity Relationship Modeling”. In: *PACIS 2012 Proceedings* (July 2012). URL: https://aisel.aisnet.org/pacis2012/152.

[35] Alvin Odon Insorio and Daniel Manansala Macandog. “Video Lessons via YouTube Channel as Mathematics Interventions in Modular Distance Learning”. In: *Contemporary Mathematics and Science Education* 3.1 (Jan. 2022). Publisher: Bastas, ep22001. ISSN: 2634-4076. DOI: 10.30935/conmaths/11468. URL: https://www.conmaths.com/article/video-lessons-via-youtube-channel-as-mathematics-interventions-in-modular-distance-learning-11468 (visited on 09/28/2024).

[36] Arbana Kadriu, Lejla Abazi Bexheti, Hyrije Abazi Alili, and Veland Ramadani. “Investigating trends in learning programming using YouTube tutorials”. en. In: *International Journal of Learning and Change* 12.2 (2020). Publisher: Inderscience Publishers, p. 190. ISSN: 1740-2875, 1740-2883. DOI: 10.1504/ijlc.2020.106721. URL: http://www.inderscience.com/link.php?id=106721 (visited on 07/12/2025).

[37] Gurjyot Singh Kalra, Ramandeep Singh Kathuria, and Amit Kumar. “YouTube Video Classification based on Title and Description Text”. In: *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. Oct. 2019, pp. 74–79. DOI: 10.1109/ICCCIS48478.2019.8974514. URL: https://ieeexplore.ieee.org/abstract/document/8974514 (visited on 07/12/2025).

[38] Zoe Kanetaki, Constantinos Stergiou, Georgios Bekas, Sébastien Jacques, Christos Troussas, Cleo Sgouropoulou, and Abdeldjalil Ouahabi. “Acquiring, Analyzing and Interpreting Knowledge Data for Sustainable Engineering Education: An Experimental Study Using YouTube”. en. In: *Electronics* 11.14 (Jan. 2022). Number: 14 Publisher: Multidisciplinary Digital Publishing Institute, p. 2210. ISSN: 2079-9292. DOI: 10.3390/electronics11142210. URL: https://www.mdpi.com/2079-9292/11/14/2210 (visited on 09/27/2024).

[39] Lena Kaul, Philipp Schrögel, and Christian Humm. “Environmental Science Communication for a Young Audience: A Case Study on the #EarthOvershootDay Campaign on YouTube”. English. In: *Frontiers in Communication* 5 (Dec. 2020). Publisher: Frontiers. ISSN: 2297-900X. DOI: 10.3389/fcomm.2020.601177. URL: https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2020.601177/full (visited on 09/27/2024).

[40] Israel Kibirige and Ronald James Odora. “Exploring the Effects of Youtube on Technology Education Students’ Cognitive Achievement in a Mechanical System Module”. In: *Perspectives in Education* 39.3 (Sept. 2021). Publisher: University of the Free State, pp. 94–108. DOI: 10.18820/2519593X/pie.v39.i3.8. URL: https://journals.co.za/doi/abs/10.18820/2519593X/pie.v39.i3.8 (visited on 09/22/2024).

[41] Cheonsoo Kim and Sung-Un Yang. “Like, comment, and share on Facebook: How each behavior differs from the other”. en. In: *Public Relations Review* 43.2 (June 2017), pp. 441–449. ISSN: 03638111. DOI: 10.1016/j.pubrev.2017.02.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S0363811116300157 (visited on 06/26/2025).

[42] Paul J Kovacs and Jeanne M Baugh. “Merging Object-Oriented Programming, Database Design, Requirements Analysis, and Web Technologies in an Active Learning Environment”. en. In: ().

[43] Fedric Kujur and Saumya Singh. “Emotions as predictor for consumer engagement in YouTube advertisement”. en. In: *Journal of Advances in Management Research* 15.2 (May 2018), pp. 184–197. ISSN: 0972-7981. DOI: 10.1108/JAMR-05-2017-0065. URL: https://www.emerald.com/insight/content/doi/10.1108/JAMR-05-2017-0065/full/html (visited on 06/26/2025).

[44] K. H. Vincent Lau, Pue Farooque, Gary Leydon, Michael L. Schwartz, R. Mark Sadler, and Jeremy J. Moeller. “Using learning analytics to evaluate a video-based lecture series”. In: *Medical Teacher* 40.1 (Jan. 2018). Publisher: Taylor & Francis, pp. 91–98. ISSN: 0142-159X. DOI: 10.1080/0142159X.2017.1395001. URL: https://www.tandfonline.com/doi/full/10.1080/0142159X.2017.1395001 (visited on 11/23/2024).

[45] Sam Lau, Sean Kross, Eugene Wu, and Philip J. Guo. “Teaching Data Science by Visualizing Data Table Transformations: Pandas Tutor for Python, Tidy Data Tutor for R, and SQL Tutor”. In: *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging education practice with education research*. DataEd ’23. New York, NY, USA: Association for Computing Machinery, June 2023, pp. 50–55. ISBN: 979-8-4007-0207-5. DOI: 10.1145/3596673.3596972. URL: https://dl.acm.org/doi/10.1145/3596673.3596972 (visited on 06/24/2025).

[46] Chung Man Lee, Eric Meyers, and Marina Milner-Bolotin. “Science Learning in YouTube Comments on Science Videos Embedding Movie References”. In: *Journal of College Science Teaching* 0.0 (). Publisher: Routledge _eprint: https://doi.org/10.1080/0047231X.2024.2389439, pp. 1–7. ISSN: 0047-231X. DOI: 10.1080/0047231X.2024.2389439. URL: https://doi.org/10.1080/0047231X.2024.2389439 (visited on 09/22/2024).

[47] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. *Towards General Text Embeddings with Multi-stage Contrastive Learning*. Aug. 2023. DOI: 10.48550/arXiv.2308.03281. URL: http://arxiv.org/abs/2308.03281 (visited on 06/24/2025).

[48] Zepei Li, Sophia Yang, Kathryn Cunningham, and Abdussalam Alawini. “Assessing Student Learning Across Various Database Query Languages”. In: *2023 IEEE Frontiers in Education Conference (FIE)*. ISSN: 2377-634X. Oct. 2023, pp. 1–9. DOI: 10.1109/FIE58773.2023.10343409. URL: https://ieeexplore.ieee.org/document/10343409 (visited on 06/24/2025).

[49] Ruben Lijo, Eduardo Quevedo, and Jose Juan Castro. “Qualitative Assessment of the Educational Use of an Electrical Engineering YouTube Channel”. In: *2023 IEEE World Engineering Education Conference (EDUNINE)*. Mar. 2023, pp. 1–6. DOI: 10.1109/EDUNINE57531.2023.10102890. URL: https://ieeexplore.ieee.org/document/10102890 (visited on 09/24/2024).

[50] Ruben Lijo, Eduardo Quevedo, Jose Juan Castro, and Ricard Horta. “Assessing Users’ Perception on the Current and Potential Educational Value of an Electrical Engineering YouTube Channel”. In: *IEEE Access* 10 (2022). Conference Name: IEEE Access, pp. 8948–8959. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3139305. URL: https://ieeexplore.ieee.org/document/9664558 (visited on 09/24/2024).

[51] Siyuan Liu, Sourav S. Bhowmick, Wanlu Zhang, Shu Wang, Wanyi Huang, and Shafiq Joty. “NEURON: Query Execution Plan Meets Natural Language Processing For Augmenting DB Education”. In: *Proceedings of the 2019 International Conference on Management of Data*. SIGMOD ’19. New York, NY, USA: Association for Computing Machinery, June 2019, pp. 1953–1956. ISBN: 978-1-4503-5643-5. DOI: 10.1145/3299869.3320213. URL: https://dl.acm.org/doi/10.1145/3299869.3320213 (visited on 06/24/2025).

[52] John C.-C. Lu. “Self-Learning Efficiency in College Virtual Course of Engineering Mathematics on YouTube”. en. In: *Engineering Proceedings* 74.1 (2024). Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 38. ISSN: 2673-4591. DOI: 10.3390/engproc2024074038. URL: https://www.mdpi.com/2673-4591/74/1/38 (visited on 09/27/2024).

[53] Francesco Maiorana. “Teaching Web Programming - An Approach Rooted in Database Principles”. In: June 2025, pp. 49–56. ISBN: 978-989-758-021-5. URL: https://www.scitepress.org/Link.aspx?doi=10.5220/0004849300490056 (visited on 06/24/2025).

[54] Eugine Tafadzwa Maziriri, Parson Gapa, and Tinashe Chuchu. “Student Perceptions towards the Use of YouTube as an Educational Tool for Learning and Tutorials”. en. In: *International Journal of Instruction* 13.2 (Apr. 2020). ERIC Number: EJ1249144, pp. 119–138. ISSN: 1694-609X. URL: https://eric.ed.gov/?id=EJ1249144 (visited on 09/22/2024).

[55] Daphne Miedema and George Fletcher. “SQLVis: Visual Query Representations for Supporting SQL Learners”. In: *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. ISSN: 1943-6106. Oct. 2021, pp. 1–9. DOI: 10.1109/VL/HCC51201.2021.9576431. URL: https://ieeexplore.ieee.org/document/9576431 (visited on 06/24/2025).

[56] Daphne Miedema, George Fletcher, and Efthimia Aivaloglou. “Expert Perspectives on Student Errors in SQL”. In: *ACM Trans. Comput. Educ.* 23.1 (Dec. 2022), 11:1–11:28. DOI: 10.1145/3551392. URL: https://dl.acm.org/doi/10.1145/3551392 (visited on 06/23/2025).

[57] Daphne Miedema, Toni Taipalus, and Efthimia Aivaloglou. “Students’ Perceptions on Engaging Database Domains and Structures”. In: *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. SIGCSE 2023. New York, NY, USA: Association for Computing Machinery, Mar. 2023, pp. 122–128. ISBN: 978-1-4503-9431-4. DOI: 10.1145/3545945.3569727. URL: https://dl.acm.org/doi/10.1145/3545945.3569727 (visited on 06/24/2025).

[58] Daphne Miedema, Toni Taipalus, Vangel V. Ajanovski, Abdussalam Alawini, Martin Goodfellow, Michael Liut, Svetlana Peltsverger, and Tiffany Young. “Data Systems Education: Curriculum Recommendations, Course Syllabi, and Industry Needs”. In: *2024 Working Group Reports on Innovation and Technology in Computer Science Education*. ITiCSE 2024. New York, NY, USA: Association for Computing Machinery, Jan. 2025, pp. 95–123. ISBN: 979-8-4007-1208-1. DOI: 10.1145/3689187.3709609. URL: https://dl.acm.org/doi/10.1145/3689187.3709609 (visited on 02/25/2025).

[59] Sriram Mohan and Aswin Punathambekar. “Localizing YouTube: Language, cultural regions, and digital platforms”. EN. In: *International Journal of Cultural Studies* 22.3 (May 2019). Publisher: SAGE Publications Ltd, pp. 317–333. ISSN: 1367-8779. DOI: 10.1177/1367877918794681. URL: https://doi.org/10.1177/1367877918794681 (visited on 07/13/2025).

[60] Miguel Ehécatl Morales-Trujillo and Gabriel Alberto García-Mireles. “Gamification and SQL: An Empirical Study on Student Performance in a Database Course”. In: *ACM Trans. Comput. Educ.* 21.1 (Dec. 2020), 3:1–3:29. DOI: 10.1145/3427597. URL: https://dl.acm.org/doi/10.1145/3427597 (visited on 06/24/2025).

[61] Lev Muchnik, Sen Pei, Lucas C. Parra, Saulo D. S. Reis, José S. Andrade Jr, Shlomo Havlin, and Hernán A. Makse. “Origins of power-law degree distribution in the heterogeneity of human activity in social networks”. en. In: *Scientific Reports* 3.1 (May 2013). Number: 1 Publisher: Nature Publishing Group, pp. 1–8. ISSN: 2045-2322. DOI: 10.1038/srep01783. URL: https://www.nature.com/articles/srep01783 (visited on 06/24/2025).

[62] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. “MTEB: Massive Text Embedding Benchmark”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2014–2037. DOI: 10.18653/v1/2023.eacl-main.148. URL: https://aclanthology.org/2023.eacl-main.148/ (visited on 06/24/2025).

[63] Ayşegül Nacak, Başak Bağlama, and Burak Demir. “Teacher Candidate Views on the Use of YouTube for Educational Purposes”. In: *Online Journal of Communication and Media Technologies* 10.2 (Mar. 2020). Publisher: Bastas, e202003. ISSN: 1986-3497. DOI: 10.29333/ojcmt/7827. URL: https://www.ojcmt.net/article/teacher-candidate-views-on-the-use-of-youtube-for-educational-purposes-7827 (visited on 09/28/2024).

[64] Wendy Ford Nina Sarkar and Christina Manzo. “To flip or not to flip: What the evidence suggests”. In: *Journal of Education for Business* 95.2 (2020), pp. 81–87. DOI: 10.1080/08832323.2019.1606771. eprint: https://doi.org/10.1080/08832323.2019.1606771. URL: https://doi.org/10.1080/08832323.2019.1606771.

[65] Daniel Pattier. “Science on Youtube: Successful Edutubers”. en. In: *TECHNO REVIEW. International Technology, Science and Society Review /Revista Internacional de Tecnología, Ciencia y Sociedad* 10.1 (Feb. 2021). Number: 1, pp. 1–15. ISSN: 2695-9933. DOI: 10.37467/gka-revtechno.v10.2696. URL: https://www.ojs.bdtopten.com/karim/index.php/revTECHNO/article/view/2696 (visited on 09/27/2024).

[66] Sophie Pavelle and Clare Wilkinson. “Into the Digital Wild: Utilizing Twitter, Instagram, YouTube, and Facebook for Effective Science and Environmental Communication”. English. In: *Frontiers in Communication* 5 (Oct. 2020). Publisher: Frontiers. ISSN: 2297-900X. DOI: 10.3389/fcomm.2020.575122. URL: https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2020.575122/full (visited on 09/27/2024).

[67] Beatriz Pérez. “Enhancing the Learning of Database Access Programming using Continuous Integration and Aspect Oriented Programming”. In: *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*. May 2021, pp. 221–230. DOI: 10.1109/ICSE-SEET52601.2021.00032. URL: https://ieeexplore.ieee.org/document/9402217 (visited on 06/24/2025).

[68] Rubén Pérez-Mercado, Antonio Balderas, Andrés Muñoz, Juan Francisco Cabrera, Manuel Palomo-Duarte, and Juan Manuel Dodero. “ChatbotSQL: Conversational agent to support relational database query language learning”. In: *SoftwareX* 22 (May 2023), p. 101346. ISSN: 2352-7110. DOI: 10.1016/j.softx.2023.101346. URL: https://www.sciencedirect.com/science/article/pii/S2352711023000420 (visited on 06/24/2025).

[69] Hidayah Rahmalan, Sharifah Sakinah Syed Ahmad, and Lilly Suriani Affendey. “Investigation on designing a fun and interactive learning approach for Database Programming subject according to students’ preferences”. en. In: *Journal of Physics: Conference Series* 1529.2 (Apr. 2020). Publisher: IOP Publishing, p. 022076. ISSN: 1742-6596. DOI: 10.1088/1742-6596/1529/2/022076. URL: https://dx.doi.org/10.1088/1742-6596/1529/2/022076 (visited on 06/24/2025).

[70] Antonio Reina, Héctor García-Ortega, Luis Felipe Hernández-Ayala, Itzel Guerrero-Ríos, Jesús Gracia-Mora, and Miguel Reina. “CADMIO: Creating and Curating an Educational YouTube Channel with Chemistry Videos”. In: *Journal of Chemical Education* 98.11 (Nov. 2021). Publisher: American Chemical Society, pp. 3593–3599. ISSN: 0021-9584. DOI: 10.1021/acs.jchemed.1c00794. URL: https://doi.org/10.1021/acs.jchemed.1c00794 (visited on 09/27/2024).

[71] Navid Rekabsaz, Mihai Lupu, and Allan Hanbury. “Exploration of a Threshold for Similarity Based on Uncertainty in Word Embedding”. en. In: *Advances in Information Retrieval*. Ed. by Joemon M Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait. Cham: Springer International Publishing, 2017, pp. 396–409. ISBN: 978-3-319-56608-5. DOI: 10.1007/978-3-319-56608-5_31.

[72] Karen Renaud and Judy van Biljon. “Teaching SQL — Which Pedagogical Horse for This Course?” en. In: *Key Technologies for Data Management*. Ed. by Howard Williams and Lachlan MacKinnon. Berlin, Heidelberg: Springer, 2004, pp. 244–256. ISBN: 978-3-540-27811-5. DOI: 10.1007/978-3-540-27811-5_22.

[73] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. “Automatic Keyword Extraction from Individual Documents”. en. In: *Text Mining*. John Wiley & Sons, Ltd, 2010, pp. 1–20. ISBN: 978-0-470-68964-6. DOI: 10.1002/9780470689646.ch1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470689646.ch1 (visited on 06/25/2025).

[74] Sonny Rosenthal. “Motivations to seek science videos on YouTube: free-choice learning in a connected society”. In: *International Journal of Science Education, Part B* 8.1 (2018), pp. 22–39. DOI: 10.1080/21548455.2017.1371357. eprint: https://doi.org/10.1080/21548455.2017.1371357. URL: https://doi.org/10.1080/21548455.2017.1371357.

[75] W. N. Sari, B. S. Samosir, N. Sahara, L. Agustina, and Y. Anita. “Learning Mathematics “Asyik” with Youtube Educative Media”. en. In: *Journal of Physics: Conference Series* 1477.2 (Mar. 2020). Publisher: IOP Publishing, p. 022012. ISSN: 1742-6596. DOI: 10.1088/1742-6596/1477/2/022012. URL: https://dx.doi.org/10.1088/1742-6596/1477/2/022012 (visited on 09/27/2024).

[76] Johannes Schildgen and Jessica Rosin. “Game-based Learning of SQL Injections”. In: *Proceedings of the 1st International Workshop on Data Systems Education*. DataEd ’22. New York, NY, USA: Association for Computing Machinery, June 2022, pp. 22–25. ISBN: 978-1-4503-9350-8. DOI: 10.1145/3531072.3535321. URL: https://dl.acm.org/doi/10.1145/3531072.3535321 (visited on 06/24/2025).

[77] Abdul Rahman Shaikh, Hamed Alhoori, and Maoyuan Sun. “YouTube and science: models for research impact”. en. In: *Scientometrics* 128.2 (Feb. 2023), pp. 933–955. ISSN: 1588-2861. DOI: 10.1007/s11192-022-04574-5. URL: https://doi.org/10.1007/s11192-022-04574-5 (visited on 09/27/2024).

[78] Shin-Shing Shin. “Teaching Method for Entity–Relationship Models Based on Semantic Network Theory”. In: *IEEE Access* 10 (2022), pp. 94908–94923. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3206028. URL: https://ieeexplore.ieee.org/document/9887961 (visited on 06/24/2025).

[79] Abdulhadi Shoufan and Fatma Mohamed. “YouTube and Education: A Scoping Review”. In: *IEEE Access* 10 (2022). Conference Name: IEEE Access, pp. 125576–125599. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3225419. URL: https://ieeexplore.ieee.org/abstract/document/9965379 (visited on 09/22/2024).

[80] Abraham Silberschatz, Henry Korth, and Sudarshan S. *Database System Concepts*. en. ISBN: 978-0-07-802215-9. URL: https://www.mheducation.com/highered/product/Database-System-Concepts-Silberschatz.html (visited on 07/04/2025).

[81] C. Snelson. “YouTube across the Disciplines: A Review of the Literature”. In: 2011. URL: https://www.semanticscholar.org/paper/YouTube-across-the-Disciplines%3A-A-Review-of-the-Snelson/f8bd63cb191a2a8da9c46a1339735b918dce072c (visited on 09/21/2024).

[82] Balaji Vasan Srinivasan, Anandhavelu Natarajan, Ritwik Sinha, Vineet Gupta, Shriram Revankar, and Balaraman Ravindran. “Will your facebook post be engaging?” In: *Proceedings of the 1st workshop on User engagement optimization*. UEO ’13. New York, NY, USA: Association for Computing Machinery, Nov. 2013, pp. 25–28. ISBN: 978-1-4503-2421-2. DOI: 10.1145/2512875.2512881. URL: https://dl.acm.org/doi/10.1145/2512875.2512881 (visited on 06/26/2025).

[83] Hector Suarez and Hooper Kincannon. “SSETGami: Secure Software Education Through Gamification”. en. In: *PROCEEDINGS ON CYBERSECURITY EDUCATION, RESEARCH AND PRACTICE* (Jan. 2017). URL: https://par.nsf.gov/biblio/10046817-ssetgami-secure-software-education-through-gamification (visited on 06/24/2025).

[84] Toni Taipalus. “Query Execution Plans and Semantic Errors: Usability and Educational Opportunities”. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA ’23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–6. ISBN: 978-1-4503-9422-2. DOI: 10.1145/3544549.3585794. URL: https://dl.acm.org/doi/10.1145/3544549.3585794 (visited on 06/24/2025).

[85] Toni Taipalus, Daphne Miedema, and Efthimia Aivaloglou. “Engaging Databases for Data Systems Education”. In: *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*. ITiCSE 2023. New York, NY, USA: Association for Computing Machinery, June 2023, pp. 334–340. ISBN: 979-8-4007-0138-2. DOI: 10.1145/3587102.3588804. URL: https://dl.acm.org/doi/10.1145/3587102.3588804 (visited on 06/23/2025).

[86] Toni Taipalus and Piia Perälä. “What to Expect and What to Focus on in SQL Query Teaching”. In: *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. SIGCSE ’19. New York, NY, USA: Association for Computing Machinery, Feb. 2019, pp. 198–203. ISBN: 978-1-4503-5890-3. DOI: 10.1145/3287324.3287359. URL: https://doi.org/10.1145/3287324.3287359 (visited on 06/23/2025).

[87] Toni Taipalus and Ville Seppänen. “SQL Education: A Systematic Mapping Study and Future Research Agenda”. In: *ACM Trans. Comput. Educ.* 20.3 (Aug. 2020), 20:1–20:33. DOI: 10.1145/3398377. URL: https://doi.org/10.1145/3398377 (visited on 06/24/2025).

[88] Jess Tan, Desmond Yeo, Rachael Neoh, Huey-Eng Chua, and Sourav S Bhowmick. “MOCHA: a tool for visualizing impact of operator choices in query execution plans for database education”. In: *Proc. VLDB Endow.* 15.12 (Aug. 2022), pp. 3602–3605. ISSN: 2150-8097. DOI: 10.14778/3554821.3554854. URL: https://dl.acm.org/doi/10.14778/3554821.3554854 (visited on 06/24/2025).

[89] Songxin Tan and Zixing Shen. “Relationship Between Cognitive Features and Social Media Engagement: An Analysis of YouTube Science Videos”. In: *IEEE Transactions on Engineering Management* 71 (2024). Conference Name: IEEE Transactions on Engineering Management, pp. 10116–10125. ISSN: 1558-0040. DOI: 10.1109/TEM.2023.3330677. URL: https://ieeexplore.ieee.org/abstract/document/10320094 (visited on 09/27/2024).

[90] The Joint Acm/Ais Is2020 Task Force, Paul Leidig, and Hannu Salmela. *A Competency Model for Undergraduate Programs in Information Systems*. en. New York, NY, USA: ACM, Jan. 2021. ISBN: 978-1-4503-8464-3. DOI: 10.1145/3460863. URL: https://dl.acm.org/doi/book/10.1145/3460863 (visited on 07/05/2025).

[91] V. Mangala Vadivu and M. Neelamalar. “Digital brand management — A study on the factors affecting customers’ engagement in Facebook pages”. In: *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*. May 2015, pp. 71–75. DOI: 10.1109/ICSTM.2015.7225392. URL: https://ieeexplore.ieee.org/document/7225392 (visited on 06/26/2025).

[92] Anikó Vágner. “Let’s learn database programming in an active way”. en. In: *Teaching Mathematics and Computer Science* 12.2 (Dec. 2014). Number: 2, pp. 213–228. ISSN: 2676-8364. DOI: 10.5485/TMCS.2014.0366. URL: https://ojs.lib.unideb.hu/tmcs/article/view/14957 (visited on 06/24/2025).

[93] Julia Valeiras-Jurado and Edgar Bernad-Mechó. “Modal density and coherence in science dissemination: Orchestrating multimodal ensembles in online TED talks and youtube science videos”. In: *Journal of English for Academic Purposes* 58 (July 2022), p. 101118. ISSN: 1475-1585. DOI: 10.1016/j.jeap.2022.101118. URL: https://www.sciencedirect.com/science/article/pii/S1475158522000388 (visited on 09/27/2024).

[94] Vivian Van Der Werf, Min Yi Zhang, Efthimia Aivaloglou, Felienne Hermans, and Marcus Specht. “Variables in Practice. An Observation of Teaching Variables in Introductory Programming MOOCs”. In: *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*. Turku Finland: ACM, June 2023, pp. 208–214. DOI: 10.1145/3587102.3588857. URL: https://dl.acm.org/doi/10.1145/3587102.3588857 (visited on 07/09/2025).

[95] Raphaela Martins Velho, Amanda Merian Freitas Mendes, and Caio Lucidius Naberezny Azevedo. “Communicating Science With YouTube Videos: How Nine Factors Relate to and Affect Video Views”. English. In: *Frontiers in Communication* 5 (Sept. 2020). Publisher: Frontiers. ISSN: 2297-900X. DOI: 10.3389/fcomm.2020.567606. URL: https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2020.567606/full (visited on 09/27/2024).

[96] Hu Wang, Hui Li, Sourav S Bhowmick, and Baochao Xu. “ARENA: Alternative Relational Query Plan Exploration for Database Education”. In: *Companion of the 2023 International Conference on Management of Data*. SIGMOD ’23. New York, NY, USA: Association for Computing Machinery, June 2023, pp. 107–110. ISBN: 978-1-4503-9507-6. DOI: 10.1145/3555041.3589713. URL: https://dl.acm.org/doi/10.1145/3555041.3589713 (visited on 06/24/2025).

[97] Ning Wang, Zachary Clowdus, Alessandra Sealander, and Robert Stern. “Geonews: timely geoscience educational YouTube videos about recent geologic events”. English. In: *Geoscience Communication* 5.2 (May 2022). Publisher: Copernicus GmbH, pp. 125–142. DOI: 10.5194/gc-5-125-2022. URL: https://gc.copernicus.org/articles/5/125/2022/ (visited on 09/27/2024).

[98] Weiguo Wang, Sourav S. Bhowmick, Hui Li, Shafiq Joty, Siyuan Liu, and Peng Chen. “Towards Enhancing Database Education: Natural Language Generation Meets Query Execution Plans”. In: *Proceedings of the 2021 International Conference on Management of Data*. SIGMOD ’21. New York, NY, USA: Association for Computing Machinery, June 2021, pp. 1933–1945. ISBN: 978-1-4503-8343-1. DOI: 10.1145/3448016.3452822. URL: https://dl.acm.org/doi/10.1145/3448016.3452822 (visited on 06/24/2025).

[99] Lizette Weilbach, Marié Hattingh, and Komla Pillay. “Using Design Patterns to Teach Conceptual Entity Relationship (ER) Data Modelling”. en. In: *Innovative Technologies and Learning*. Ed. by Yueh-Min Huang, Chin-Feng Lai, and Tânia Rocha. Cham: Springer International Publishing, 2021, pp. 228–238. ISBN: 978-3-030-91540-7. DOI: 10.1007/978-3-030-91540-7_25.

[100] Sophia Yang, Zepei Li, Geoffrey L. Herman, Kathryn Cunningham, and Abdussalam Alawini. “Uncovering Patterns of SQL Errors in Student Assignments: A Comparative Analysis of Different Assignment Types”. In: *2023 IEEE Frontiers in Education Conference (FIE)*. ISSN: 2377-634X. Oct. 2023, pp. 01–09. DOI: 10.1109/FIE58773.2023.10343207. URL: https://ieeexplore.ieee.org/document/10343207 (visited on 06/24/2025).

[101] Shupei Yuan and Hang Lu. “Examining a conceptual framework of aggressive and humorous styles in science YouTube videos about climate change and vaccination”. en. In: *Public Understanding of Science* 31.7 (Oct. 2022). Publisher: SAGE Publications Ltd, pp. 921–939. ISSN: 0963-6625. DOI: `10.1177/09636625221091490`. URL: `https://doi.org/10.1177/09636625221091490` (visited on 09/27/2024).

A

Prompts

A.1. Prompt for GPT-based Relevance Classification

Video Relevance Classification Prompt

Given the following YouTube video INFORMATION, we are looking to see if it matches our KEYWORD LIST. Reply with “1” if and only if INFORMATION is an “instructional video” on any data system topic that matches KEYWORD LIST. Otherwise, reply “0”.
Instructional Video Definition: A video is instructional if it is designed to educate, train, or inform viewers by demonstrating a process, explaining a concept, or providing expert insights.
Exclusions: Do not consider news reports, marketing/promotional material, or legal interpretations/explanations.
—- INFORMATION START —–
Title: {title}
Description: {description}
Video Transcript: {transcript}
—- INFORMATION END —–
—- KEYWORD LIST START —–
[... all search queries ...]
—- KEYWORD LIST END —–

A.2. Prompt for Grouping SQL Index Terms

SQL Index Term Grouping Prompt

You are a linguistic and domain expert. Given the list of technical terms below (mostly related to SQL and databases), group them into coherent, non-overlapping subgroups. Each group should contain terms that refer to the same concept, variant forms, plural/singular differences, or phrases commonly used together in the same context.
Please follow these rules:

- Group terms by meaning, not just word similarity (e.g., `aggregation`, `aggregation in sql`, and `aggregation operation` go together).
- Include synonyms, plurals/singulars, and closely related forms in the same group.
- Prefer 1–5 words per group label.
- The labels should be really specific, with small granularity.
- Use bullet points, where each group starts with a bolded group name, followed by a list of related terms.
- Do not skip or omit any term. Every term in the list must appear in exactly one group.

Here is the list:
[raw list of all index terms from the three textbooks]

A.3. Prompt for LLM-based SQL Subtopic Classification

Sample Prompt for SQL Subtopic Classification

You are an expert in SQL topic classification. Given a transcript snippet of a SQL tutorial video, your task is to determine which of the predefined SQL subcategories the content belongs to.

Instructions:

- You may use the presence or semantic meaning of keywords as soft signals to guide classification.
- Only include subcategories that are directly relevant. Avoid selecting broad or tangential categories unless clearly supported by the transcript.
- If no match is confidently inferred from either explicit terms or implicit context, return an empty list.
- Your response must only include the JSON output enclosed in triple backticks.

========== Example 1 ==========

Title: SQL Constraints Tutorial - Enforcing Data Rules

Transcript Snippet: Let's say you want to ensure that every employee in your table has a non-null salary value. You can use the NOT NULL constraint on the salary column when defining the table.

Output:

```
{
  "matching_categories": [
    "Domain & Check Constraints"
  ]
}
```

========== Example 2 ==========

Title: How to Install MySQL on Windows 11

Transcript Snippet: First, go to the MySQL website, download the installer, and follow the steps to install the MySQL Workbench and server on your machine.

Output:

```
{
  "matching_categories": []
}
```

========== Example 3 ==========

Title: Filtering and Sorting Data in SQL

Transcript Snippet: We'll use the WHERE clause to limit the rows, then sort the result with ORDER BY. You can also use AND and OR to combine multiple filter conditions.

Output:

```
{
  "matching_categories": [
    "Expressions & Syntax",
    "Logical Connectives",
    "Ordering & Limits"
  ]
}
```

========== SQL Category Keywords ==========

[mapping of SQL subtopics to textbook-derived keyword terms as detailed in Appendix B.2]

========== Classification Task ==========

Title: {title}

Transcript Snippet: {transcript snippet truncated at 3,000 tokens}

Output:

B

Textbook Derived SQL Index Terms and Grouped Subtopics

B.1. Textbook SQL Index Terms

SQL index terms referenced in *Database Management Systems* (3rd Edition) by Ramakrishnan and Gehrke, Chapter 5 on SQL: Queries, Constraints, Triggers:

- Active databases
- Aggregation in SQL
- Assertions in SQL
- AVG
- Collations in SQL
- Conceptual evaluation strategy
- Correlated queries
- COUNT
- CREATE DOMAIN
- CREATE TRIGGER
- CREATE TYPE
- Data Definition Language (DDL)
- Data Manipulation Language (DML)
- Dates and times in SQL
- Difference operation
- Distinct type in SQL
- Domain constraints
- Duplicates in SQL
- Events activating triggers
- Expressions in SQL
- Grouping in SQL
- IBM DB2
- Informix UDS
- Intersection operation
- MAX
- MIN
- Multisets
- Nested queries
- Outer joins
- Packages in SQL1999
- Row-level triggers
- Set comparisons in SQL
- Set operators
- SQL
- Statement-level triggers
- Strings in SQL
- SUM
- Triggers
- Union operation

SQL index terms referenced in *Database System Concepts* (7th Edition) by Silberschatz, Korth, and Sudarshan, Part 1: Relational Languages:

- ADO.NET
- ANSI (American National Standards Institute)
- Boolean operations
- C
- C++
- CLI (Call Level Interface) standards
- CLR (Common Language Runtime)
- Call Level Interface (CLI) standards
- Cartesian products
- Cartesian-product operation
- Common Language Runtime (CLR)
- DriverManager class
- EXEC SQL
- IBM DB2
- ISO (International Organization for Standardization)
- International Organization for Standardization (ISO)
- JDBC (Java Database Connectivity)
- Java
- LINQ (Language Integrated Query)
- Language Integrated Query (LINQ)
- Microsoft SQL Server
- MySQL
- ODBC (Open Database Connectivity)
- Oracle
- PL/SQL
- PSM (Persistent Storage Module)
- Perl
- Persistent Storage Module (PSM)
- PostgreSQL
- Python
- ResultSet object
- SQL (Structured Query Language)
- SQL environment
- SQL injection
- Sequel
- Statement object
- System R
- Tcl
- TransactSQL
- United States
- Unix
- VPD (Virtual Private Database)

Virtual Private Database (VPD)
Visual Basic
add constraint
advanced SQL
after triggers
aggregate functions
aggregation
aggregation operation
aliases
all construct
alter table
alter trigger
alter type
and connective
and operation
anti-join operation
application program interfaces (APIs)
arity
as clause
as of period for
asc expression
assertions
assignment operation
associations
atomic domains
atomicity
attributes
authorization
authorization graph
automatic commit
avg expression
backup
banking
base query
before triggers
begin atomic...end
binary operations
blobs
bugs
business logic
call statement
callable statements
candidate keys
cascades
case construct
cast
catalogs
change relation
char
check clause
check constraints
clobs
coalesce function
commit work
compatible relations
conformance levels
conversions
correlated subqueries
correlation name
correlation variables
count function
count values
create assertion
create distinct type
create domain
create function
create index
create procedure
create recursive view
create role
create schema
create sequence construct
create table...as
create table...like
create temporary table
create type
create unique index
create view
cross join
cross-tabulation
cube construct
current date
data-definition language (DDL)
data-manipulation language (DML)
database instance
database-management systems (DBMSs)
databases
databases administrator (DBA)
datetime data type
debugging
declarative queries
declare statement
decode
default values
deferred integrity constraints
deletion
delta relation
desc expression
disable trigger
distinct types
domain constraints
domain of attributes
drop index
drop schema
drop table
drop trigger
drop type
dynamic SQL
embedded SQL
embedded databases
empty relations test
equivalence
equivalent queries
escape
every function
except all
except clause
except construct
except operation
exception conditions
exceptions
execute privilege
exists construct
external language routines
false values
fetching
fixed point of recursive view definition
float
for each row clause
for each statement clause
foreign keys
from clause
full outer joins
functional dependencies
functional query language
functions
getColumnCount method
getConnection method
getFloat method
getString method
grant command
grant privileges
granted by current role
group by clause
grouping sets construct
handlers
having clause
hierarchies
histograms
host language
identity specification
if clauses
if-then-else statements
imperative query language
in construct
initially deferred integrity constraints
inner joins
insertion
instead of feature
integrity constraints
intermediate SQL
intersect all
intersect operation
interval data type
is not null
is not unknown
is null
is unknown
iteration
join conditions
join operation
join using operation
joins
keys
language constructs
large-object types
lateral clause
left outer join
like operator
limit clause

localtimestamp
materialized views
max function
metadata
min function
minus
monotonic queries
multiset except
multiset relational algebra
natural join operation
natural joins
nested subqueries
next method
nondeclarative actions
not connective
not exists construct
not in construct
not null
not operation
not unique construct
null values
numeric
nvarchar
on condition
on delete cascade
on update cascade
or connective
or operation
order by clause
outer union operation
outer-join operation
parameter style general
parameterized views
partitions
passwords
period declaration
pivot attribute
pivot clause
pivot-table
pivoting
pointers
prepared statements
primary keys
privileges
procedural languages
procedures
programming languages
project operation
queries
query languages
ranking
real, double precision
recursive queries
referenced relation
references
references privilege
referencing new row as clause
referencing new table as clause
referencing old row as clause
referencing old table as clause
referencing relation
referential integrity
relation
relational algebra
relational instance
relational model
relational schema
relational-algebra expressions
rename operation
repeat loop
repeat statements
restriction
revoke privileges
right outer join
roles
rollback
rollback work
rollup clause
rollup construct
row-level authorization
sandbox
scalar subqueries
schema diagrams
schemas
security
select all
select authorization, privileges and
select clause
select distinct
select operation
select privilege
select-from-where
semijoin operation
set autocommit off
set clause
set default
set null
set operations
set role
set statement
set-difference operation
some construct
some function
sql security invoker
sqlstate
standards
start with/connect by prior syntax
string operations
structured types
sum function
superkeys
superusers
syntax
sys.context function
table alias
table functions
tables
temporal validity
terms
then clause
timestamps
timezone
transaction control
transition tables
transition variables
transitive closure
triggers
trim
true predicate
true values
try-with-resources construct
tuple variables
tuples
types
unary operations
union all
union of sets
union operation
unique construct
unique key values
university database
unknown values
updatable result sets
updates
user-defined types
valid time
varchar
versions period for
view definition
view maintenance
views
when clause
when statement
where clause
while loop
while statements
windows and windowing
with check option
with clause
with data clause
with grant option
with recursive clause
with timezone specification

SQL index terms referenced in *Database Systems: The Complete Book* by Garcia-Molina, Ullman, and Widom, Chapter 6: The Database Language SQL:

Aggregation
ALL
And

ANSI
ANY
AS
Astrohan, M. M.
Atomicity
Attribute
Average
Berenson, H.
Bernstein, P. A.
Bit string
Case sensitivity
Celko, J.
Chamberlin, D. D.
Commit
Corrolated subquery
Count
CROSS JOIN
Darwen, H.
Date
Date, C. J.
Deletion
Difference
Dirty data
Duplicate elimination
Escape character
EXISTS
FROM
Generic interface
Gray, J. N.
GROUP BY
Gulutzan, P.
HAVING
Host language
IN
Insertion
Intersection
Isolation level
Join
Left outerjoin
Lexicographic order
LIKE
Maximum
Melton, J.
Minimum
Natural join
Negation
Null value
O’Neil, E.
O’Neil, P.
Or
ORDER BY
Outerjoin
Pelzer, P.
Product
Projection
Read commited
Read uncommited
Read-only transaction
Relational algebra
Repeatable read
Right outerjoin
Rollback
SELECT
Selection
Serializability
Set
Simon, A. R.
SQL
Subquery
Sum
System R
Three-valued logic
Time
Timestamp
Transaction
Truth value
Union
UNKNOWN
Update
WHERE

B.2. Mapping of SQL Subtopics to Textbook-Derived Keyword Terms

Subtopic	**Keyword Terms**
Active Databases	Active databases
Aggregate Functions	Aggregation, AVG, COUNT, MAX, MIN, SUM
Aliases & Correlation	aliases, table alias, correlation name, correlation variables, tuple variables, lateral clause, AS
Arity	arity
Atomicity & Domains	atomic domains, atomicity
Authorization & Privileges	authorization, authorization graph, privileges, grant command, revoke privileges, select privilege, references privilege, roles, create role, set role, row-level authorization, sql security invoker, passwords, security, sys.context function, superusers, VPD (Virtual Private Database), granted by current role, execute privilege
Backup & Recovery	backup
Business Logic	business logic
Common Language Runtime (CLR)	Common Language Runtime (CLR)
Cartesian & Product	Cartesian products, Product
Catalogs & Metadata	catalogs
Change Tracking & Delta	delta relation, change relation
Cursor Operations	fetching, updatable result sets, next method
Data Definition Language (DDL)	Data definition language (DDL)
Data Manipulation Language (DML)	Data Manipulation Language (DML), Insertion, deletion, Update, change relation, tuples
Data Types - Large Objects	large-object types, blobs, clobs

Continued on next page

Continued from previous page

Subtopic	**Keyword Terms**
Data Types - Scalar	char, varchar, nvarchar, numeric, float, real, double precision, Bit string, datetime data type, timestamp, interval data type
Database Systems	IBM DB2, Microsoft SQL Server, MySQL, Oracle, PostgreSQL, Informix UDS, System R, database-management systems (DBMSs), database instance, databases, databases administrator (DBA)
Difference & EXCEPT	minus, except all, except clause, except construct, Difference operation, set-difference operation
Dirty Data	Dirty data
Domain & Check Constraints	domain constraints, check constraints, check clause, default values, set default, not null, Assertions in SQL, create assertion, add constraint, CREATE DOMAIN, domain of attributes
Duplicate Handling	Duplicate elimination, Duplicates in SQL
Embedded & Dynamic SQL	embedded SQL, embedded databases, dynamic SQL, EXEC SQL, host language
Example Databases	banking, university database, sandbox
Exceptions & Debugging	exceptions, exception conditions, sqlstate, debugging, bugs
Expressions & Syntax	Expressions in SQL, syntax, WHERE, FROM, IN, in construct, not in construct, not exists construct, some construct, some function, EXISTS, ANY, ALL, case construct, decode, empty relations test
Fetch/Result APIs	application program interfaces (APIs), Call Level Interface (CLI) standards, Open Database Connectivity (ODBC), Generic interface, DriverManager class, getConnection method, Statement object, ResultSet object, getFloat method, getString method, ADO.NET, try-with-resources construct, jdbc (java database connectivity), getcolumncount method
Group BY & Having	GROUP BY, group by clause, Grouping in SQL, grouping sets construct, rollup clause, rollup construct, HAVING, cube construct
Hierarchies	hierarchies, start with/connect by prior syntax
Identity Columns	identity specification
Index	create index, create unique index, drop index
Integrity Constraints	integrity constraints, deferred integrity constraints, initially deferred integrity constraints, set null
Join Operations	Join, Natural join, CROSS JOIN, Left outerjoin, Right outerjoin, inner joins, Outer join, anti-join operation, semijoin operation, on condition, join using operation, full outer join
Key Constraints	keys, candidate keys, primary keys, superkeys, unique construct, unique key values, not unique construct
Language Integrated Query (LINQ)	Language Integrated Query (LINQ)
Logical Connectives	and connective, or connective, not connective, not operation, Negation, Boolean operations, or operation
Null & Unknown Handling	Null value, UNKNOWN, unknown values, is null, is not null, is unknown, is not unknown, Three-valued logic, Truth value, true predicate, true values, false values
Operating Systems	Unix
Ordering & Limits	ORDER BY, asc expression, desc expression, limit clause, Lexicographic order
Partitioning	partitions
Pointers	pointers

Continued on next page

Continued from previous page

Subtopic	**Keyword Terms**
Prepared Statements	prepared statements, parameter style general, call statement, parameterized views
Procedures & PSM	procedures, create procedure, functions, create function, handlers, procedural languages, Persistent Storage Module (PSM), PL/SQL, begin atomic...end, repeat loop, repeat statements, while loop, while statements, if clauses, if-then-else statements, then clause, when clause, when statement, nondeclarative actions, Packages in SQL: 1999, declare statement, iteration, external language routines
Programming Languages	C, C++, Java, Perl, Python, Tcl, Visual Basic, TransactSQL, programming languages
Projection & Project Operation	project operation, Projection, Attribute
Queries & Paradigms	queries, query languages, declarative queries, functional query language, imperative query language
Recursive Queries	recursive queries, with recursive clause, fixed point of recursive view definition, transitive closure
Referential Integrity	referential integrity, references, referenced relation, referencing relation, referencing new row as clause, referencing new table as clause, referencing old row as clause, referencing old table as clause, on delete cascade, on update cascade, cascades, foreign keys
Relational Model & Algebra	relation, relational model, relational schema, relational instance, relational algebra, relational-algebra expressions, functional dependencies, multiset relational algebra, Multisets, multiset except, Set comparisons in SQL, compatible relations, equivalence, equivalent queries, Conceptual evaluation strategy, monotonic queries, binary operations, unary operations, rename operation
Row-Level Security	Row-level triggers
SQL Standards & History	Structured Query Language (SQL), Sequel, American National Standards Institute (ANSI), International Organization for Standardization (ISO), standards, SQL environment, conformance levels
Scalar Functions	cast, coalesce function, every function
Schema	create schema, drop schema, schemas, schema diagrams
Security	SQL injection
Select Variants	SELECT, select clause, select distinct, select all, select operation, select privilege, select authorization, privileges and, select-from-where, Selection, base query, restriction
Sequence	create sequence construct
Set & Assignment	set clause, Set, assignment operation, set null, set statement
Set Operations	Union, union all, union of sets, intersect all, Intersection, outer union operation, set operations, Set operators
Statistics	histograms
String Functions	Strings in SQL, string operations, trim, LIKE, escape, Escape character, Case sensitivity, Collations in SQL
Subqueries	Subquery, Nested queries, nested subqueries, correlated subqueries, scalar subqueries
Table	create table...as, create table...like, create temporary table, alter table, drop table, tables
Table Functions	table functions

Continued on next page

Continued from previous page

Subtopic	**Keyword Terms**
Temporal Concepts	as of period for, versions period for, period declaration, valid time, temporal validity, current date, localtimestamp, Dates and times in SQL, timezone, timestamp
Transactions & Isolation	Transaction, Commit, rollback, rollback work, automatic commit, set autocommit off, transaction control, Read commited, Read uncommited, Repeatable read, Serializability, Isolation level, Read-only transaction
Triggers	triggers, after triggers, before triggers, Row-level triggers, Statement-level triggers, Events activating triggers, CREATE TRIGGER, alter trigger, disable trigger, drop trigger, for each row clause, for each statement clause, instead of feature, transition tables, transition variables
Type	distinct type, create distinct type, CREATE TYPE, alter type, drop type, types, user-defined types, structured types
View	views, create view, parameterized views, create recursive view, view definition, materialized views, view maintenance
WITH Clauses	with clause, with data clause, with check option, with grant option, with timezone specification
Windowing & Pivoting	windows and windowing, pivot clause, pivot attribute, pivot-table, pivoting, ranking

C

SQL Subtopics Classification Samples and Statistics

C.1. Manual Validation Sample for SQL Subtopics Classification

Evaluation Sample (JSONL Format)

Sample ID: sample_029

Passage:

In general, a cross-tab is a table derived from a relation (say, R), where values for some attribute of relation R (say, A) become attribute names in the result; the attribute A is the pivot attribute. Cross-tabs are widely used for data analysis, and are discussed in more detail in Section 11.3.

Several SQL implementations, such as Microsoft SQL Server, and Oracle, support a pivot clause that allows creation of cross-tabs. Given the sales relation from Figure 5.17, the query:

```
select * from sales pivot (sum(quantity) for color in ('dark', 'pastel',
'white'))
```

returns the result shown in Figure 5.18.

Note that the for clause within the pivot clause specifies (i) a pivot attribute (color, in the above query), (ii) the values of that attribute that should appear as attribute names in the pivot result (dark, pastel and white, in the above query), and (iii) the aggregate function that should be used to compute the value of the new attributes (aggregate function sum, on the attribute quantity, in the above query).

The attribute color and quantity do not appear in the result, but all other attributes are retained. In case more than one tuple contributes values to a given cell, the aggregate operation within the pivot clause specifies how the values should be combined. In the above example, the quantity values are aggregated using the sum function.

A query using pivot can be written using basic SQL constructs, without using the pivot construct, but the construct simplifies the task of writing such queries.

Index Terms: pivot clause, pivot attribute, pivoting, pivot-table

Ground Truth: Windowing & Pivoting

Qwen3 Prediction: Windowing & Pivoting

LLaMA3 Prediction: Expressions & Syntax, Join Operations, Set Operations, Windowing & Pivoting

C.2. Representative Samples for SQL Subtopic Classification Results

Sample 1: Video 5OpBjU-OWh8

Video Title: How to Join two or more than two Tables using multiple columns | How to Join Multiple Tables #Joins

Transcript Snippet:

> Hey everyone, welcome to Data Millennials. I am Atul and in this video we are going to discuss about joining multiple tables. Suppose you have three tables: student data, student course data, and student marks data. In your student data table, you have details about students. In your student course data, you have details about the courses in which students have enrolled. The student marks data contains marks for the corresponding subjects or courses. Now, if you have to join all of these three tables, how can you join them? Let's go to our SQL workbench and first run all three SELECT queries to see the data in our SQL tables. First, SELECT FROM student_data returns 10 records. Then, SELECT FROM student_course_data shows 60 records because each student is enrolled in five different courses—so there's duplicacy in roll numbers, but the records are unique at roll number and course name level. SELECT FROM student_marks_data also returns 60 rows, with columns roll number, course name, and marks. This table contains marks for each course for every student.
>
> What we have to do is get the name, class, and roll number from student data, the courses from student course data, and the marks from student marks data. We'll consider student data as the first (left) table: SELECT sd.roll_number, sd.class, sd.name FROM student_data sd. Now we left join this with student_course_data: LEFT JOIN student_course_data scd ON sd.roll_number = scd.roll_number, and select scd.course_name. Running this query returns 50 rows because student data is the left table...

Note: This transcript snippet is truncated for brevity; full input (up to 3000 tokens) was provided to the model during classification.

Qwen3-8B Predicted Categories: Join Operations

Sample 2: Video AZ29DXaJ1Ts

Video Title: SQL Project | SQL Case Study to SOLVE and PRACTICE SQL Queries | 20+ SQL Problems

Transcript Snippet:

> Hey everyone. In this video let's work on an SQL case study. As part of this case study, first we will try to download the data set from Kaggle. Once we have the data set, we'll then try to upload it into our database using a very simple Python script. Once we have the data available, we'll try to analyze it and then we will try to solve around 20 plus SQL queries as part of this SQL case study. Now you can call this like a case study or you can call it like an SQL project, but basically, if you want to solve basic to intermediate level of SQL queries, then this is pretty perfect. Now, of course, I will not be able to solve all the 20 plus SQL queries as part of this video, because then the video is going to be very long, but rather I'll take a handful of the queries and I'll try to solve it during this video. But for all the remaining SQL queries, the data set, the scripts and everything else—you'll find it in my blog. I'll leave the link in the video description. Let's start. So first of all, let's take our data set from Kaggle. I will leave this link in the description so you can go through this link and can download the data set yourself. The name is Famous Paintings. It's given by Maxwell and I think he has taken this data set from Data World. He has mentioned that detail there. Now, if I go down, you can see that this data set has eight different files and it has information about artist, paintings, museums, etc. The first thing that we will do is click on the download so that all these eight CSV files are downloaded into our system. I have already downloaded them and I have placed them in one of my folders. The next thing that we need to do is load this data into our database. Now, there are two ways I can do this. One is I can go into my database. I'm using PostgreSQL database and the PGAdmin tool. Of course, you can use any other database of your choice...

Note: This transcript snippet is truncated for brevity; the full input (up to 3000 tokens) was provided to the model during classification.

Qwen3-8B Predicted Categories: Data Loading & Integration, Python Integration, Data Import/Export

C.3. Full SQL Subtopic Coverage Statistics

Table C.1: SQL subtopics covered in the analyzed YouTube dataset ($N = 4{,}242$ SQL-related videos with transcripts), showing video count, relative frequency, and median engagement metrics and video duration.

Subtopic	Videos (%)	Median View	Median Like	Median Comment	Median Duration (min)
Data Manipulation Language (DML)	1,335 (31.47%)	5,011	78.5	5	8.75
Expressions & Syntax	882 (20.79%)	4,883	101.5	5	12.12
Join Operations	717 (16.90%)	8,509	188	10	12.02
Group BY & Having	682 (16.08%)	1,986	35.5	3	10.46
Subqueries	650 (15.32%)	5,481	89	6	11.67
Aggregate Functions	582 (13.72%)	666.5	14	1	10.20
Ordering & Limits	410 (9.66%)	3,669.5	90.5	4	12.72
Data Definition Language (DDL)	326 (7.69%)	64,572	1,176	56	35.54
Select Variants	303 (7.14%)	4,828	52	4	7.47
Logical Connectives	268 (6.32%)	9,998	219	12	14.58
Transactions & Isolation	265 (6.25%)	7,933	157	12	12.38
Queries & Paradigms	224 (5.28%)	48,525.5	1,139.5	59	44.59
Windowing & Pivoting	193 (4.55%)	5,423	163	9.5	15.83
Database Systems	150 (3.54%)	75,692	1,283	58.5	41.41
Relational Model & Algebra	134 (3.16%)	72,926	1,862	64.5	52.98
Data Types - Scalar	129 (3.04%)	54,774	979	56	46.75
Table	128 (3.02%)	16,551	270	18	13.06
Schema	96 (2.26%)	52,801	1,106	49	52.05
Duplicate Handling	95 (2.24%)	3,178	69	5	6.45
Recursive Queries	90 (2.12%)	1,103.5	21.5	3	14.59
String Functions	88 (2.07%)	5,339.5	154	10	14.84
WITH Clauses	74 (1.74%)	1,939.0	41	4	7.84
Procedures & PSM	71 (1.67%)	10,326	197	6	11.32
Key Constraints	62 (1.46%)	44,591.5	954	44.5	17.36
Index	61 (1.44%)	12,105	368	20	25.38
Domain & Check Constraints	57 (1.34%)	32,981	488	23	38.15
Integrity Constraints	53 (1.25%)	34,772	337	22	26.03
Scalar Functions	53 (1.25%)	8,677	154	6	19.70
Referential Integrity	50 (1.18%)	18,077.5	309	17	12.81
Security	48 (1.13%)	40,855	454.5	22	35.12
SQL Standards & History	46 (1.08%)	53,824.5	1,222.5	32	17.67
Null & Unknown Handling	44 (1.04%)	4,881.5	154	3	19.70
Triggers	42 (0.99%)	4,828.5	94	3	14.64
Data Types - Large Objects	39 (0.92%)	30,773	604	42	96.00
Set Operations	36 (0.85%)	4,188.5	176	4	21.37
Backup & Recovery	27 (0.64%)	12,105	323	22	21.60
Aliases & Correlation	22 (0.52%)	2,927.5	62.5	2.5	11.95
Temporal Concepts	13 (0.31%)	8,271	169	4	7.52
View	12 (0.28%)	2,567.5	191	1	11.26

Continued on next page

Continued from previous page

Subtopic	**Videos (%)**	**Median View**	**Median Like**	**Median Comment**	**Median Duration (min)**
Programming Languages	12 (0.28%)	35,816.5	412	19.5	15.83
Authorization & Privileges	10 (0.24%)	64,167	778	26.5	20.58
Projection & Project Operation	9 (0.21%)	4,911	74	2	20.98
Cartesian & Product	8 (0.19%)	3,276.5	154	1.5	10.75
Statistics	8 (0.19%)	59,897.5	1,426.5	71	10.12
Hierarchies	8 (0.19%)	1,705	28	3	9.95
Identity Columns	8 (0.19%)	1,455.5	27.5	0.5	5.33
Embedded & Dynamic SQL	7 (0.17%)	41,735	558	22	17.78
Prepared Statements	6 (0.14%)	24,665.5	315.5	14	16.23
Partitioning	5 (0.12%)	217	7	0	19.70
Catalogs & Metadata	5 (0.12%)	53,420	1,162	138	136.92
Cursor Operations	5 (0.12%)	41,735	558	22	21.60
Table Functions	4 (0.09%)	22,248.5	271	3	6.95
Type	4 (0.09%)	23,090	374.5	18	30.20
Set & Assignment	3 (0.07%)	2,540	154	3	715.48
Active Databases	3 (0.07%)	2,540	154	3	715.48
Business Logic	3 (0.07%)	567	15	0	1.02
Fetch/Result APIs	3 (0.07%)	41,735	558	22	21.60
Dirty Data	2 (0.05%)	32,693	1,110	69	72.61
Common Language Runtime (CLR)	2 (0.05%)	1,518	15.5	3.5	12.07
Sequence	1 (0.02%)	49,784	787	31	17.92
Language Integrated Query (LINQ)	1 (0.02%)	60,855	161	15	4.57