Evaluating Semantic Search Versus Keyword Search For Educational Video Retrieval in Technical Subjects

Aiden Christian

Abstract

The rapid expansion of online educational video platforms has transformed how learners engage with technical subjects, yet the search systems that govern content discovery have received comparatively little scrutiny. Traditional keyword-based retrieval depends on exact lexical overlap, systematically failing learners who lack the precise domain vocabulary needed to articulate effective queries—a barrier that falls disproportionately on novices. This paper evaluates whether semantic search, powered by transformer-based language models and vector similarity search, offers a more learner-centered alternative for educational video retrieval in technical subjects. Two retrieval systems—a keyword-based baseline and a semantic search system using dense embeddings stored via pgvector—were implemented over the same dataset of publicly available technical educational videos and evaluated across thirty queries spanning three categories: exact technical terms, natural language questions, and paraphrased conceptual expressions. Retrieval alignment was assessed qualitatively, with illustrative frequency data reported to characterize observed patterns. Results indicate that semantic search demonstrated substantially higher intent alignment for natural language (~85% vs. ~40%) and paraphrased queries (~80% vs. ~35%), while both systems performed comparably on exact technical terms (~92% vs. ~90%). These findings suggest that the choice of retrieval architecture in educational platforms is not merely a technical decision but a pedagogical one, with direct implications for learner equity, content accessibility, and the design of AI-driven educational systems. Key methodological limitations include single-researcher relevance judgment without inter-rater reliability checks, the absence of an annotated gold standard for optimal retrieval, and a non-parallel query design that may conflate query construction effects with retrieval approach effects.

1. Introduction

Every search is an act of translation. The learner holds a question—often vague, partially formed, colored by what they do not yet know—and must convert it into words precise enough for a machine to act on. In most educational video platforms — particularly the broad ecosystem of LMS and open courseware platforms beyond the largest players — that machine remains a keyword-based search engine: a system that retrieves content by matching the exact tokens in a query against an indexed corpus. While platforms like YouTube have begun integrating NLP-based content understanding into discovery systems [21], this shift has been uneven, and the explicit search experience on dedicated educational platforms has lagged considerably behind. The translation burden falls entirely on the learner. For expert users who command the vocabulary of a domain, this works well. For novices navigating unfamiliar territory, it fails in exactly the moments when effective guidance matters most. This is not a limitation unique to video search — it is a structural property of keyword retrieval in any domain — but it is

especially consequential in educational contexts, where the platform's core purpose is to serve users whose vocabulary is precisely what they are trying to develop.

This asymmetry is not incidental. It is structural. Keyword-based retrieval treats language as a collection of discrete, interchangeable tokens and measures relevance by their frequency and distribution [14]. A student who types “why my code gets slower with more loops” receives different results than one who types “time complexity”—even though both are seeking the same concept. The system cannot resolve this gap because it has no model of meaning, only a model of words. In technical educational contexts, where learners’ understanding of a domain is necessarily incomplete, this design choice consistently disadvantages the people the platform is meant to serve.

Empirical research on educational search behavior bears this out. [2] found that while 94% of students report using search engines for learning, only 26% are rated as proficient at formulating effective search queries, with vocabulary mismatch accounting for approximately 40% of irrelevant results. [19] identified vocabulary mismatch as a primary obstacle to effective retrieval in educational information systems. These findings converge on a troubling conclusion: keyword-based search is most reliable for users who need it least.

Semantic search offers a structural alternative. Rather than matching tokens, semantic retrieval systems encode both queries and documents as dense vector representations—embeddings—that capture contextual meaning in a high dimensional space [20] - [10]. A query expressed as “why my code gets slower with more loops” and a document discussing time complexity and algorithmic efficiency can be recognized as conceptually proximate even when they share no surface vocabulary. This capability is precisely what educational retrieval demands.

This paper evaluates whether semantic search outperforms keyword search for educational video retrieval in technical subjects. To answer this question, two retrieval systems were implemented over the same dataset of publicly available technical educational videos: a keyword-based baseline and a semantic search system using transformer-based embeddings stored via PostgreSQL with the pgvector extension [1]. Both systems were evaluated across thirty queries spanning three categories—exact technical terms, natural language questions, and paraphrased conceptual expressions—with retrieval outcomes assessed qualitatively and summarized with illustrative frequency data.

The paper proceeds as follows. Section 2 reviews relevant literature on keyword-based retrieval, the emergence of semantic search, embedding models, and their application to educational and multimedia contexts. Section 3 describes the methodology and system architecture in detail. Section 4 reports results by query category. Section 5 discusses implications, including scalability and interpretability challenges that complicate real-world deployment. Section 6 addresses limitations and proposed solutions, and Section 7 concludes with directions for future research.

A note on scope: this evaluation uses qualitative assessment of retrieval alignment as its primary method, supplemented by illustrative frequency estimates derived from the evaluation

set. This approach prioritizes ecological validity and instructional relevance over statistical generalization. Its limitations are addressed directly in Sections 3.6 and 6. The ambition of this paper is not to resolve the keyword-versus-semantic debate definitively, but to sharpen it—by situating it in the specific context of learner-centered educational retrieval and making explicit the equity stakes that are too often left implicit.

2. Literature Review

2.1 Foundations of Keyword-Based Information Retrieval

Keyword-based search has served as the operational backbone of information retrieval for decades. The foundational mechanisms—inverted indexes, term frequency–inverse document frequency (TF–IDF), and BM25—are well understood, computationally efficient, and remarkably effective for users who know what to ask for [14] - [3]. Documents are scored according to how frequently query terms appear and how distinctively they identify a document within the corpus. The result is a system that rewards lexical precision and punishes imprecision.

This architecture carries a built-in assumption: that effective users can bridge the gap between their mental model of a concept and the vocabulary used to describe it in indexed content. In professional search contexts—legal discovery, medical literature review, enterprise knowledge management—this assumption is often reasonable. Users are domain experts who have internalized the relevant vocabulary through years of practice. In educational contexts, the assumption fails by design. Learners are, by definition, in the process of acquiring vocabulary they do not yet possess. A system that requires domain vocabulary as its primary input cannot effectively serve users whose most pressing need is to develop that vocabulary.

Research has consistently documented this failure mode. [19] identify vocabulary mismatch as a central obstacle in educational video retrieval, noting that learners' informal query language rarely aligns with the indexed vocabulary of instructional content. [9] situate this problem within the broader challenge of lexical semantics in natural language processing: words are ambiguous, synonymous, and context-dependent in ways that bag-of-words models cannot resolve. The vocabulary mismatch problem is not a bug in keyword search—it is a fundamental consequence of its architecture.

2.2 The Emergence of Semantic Search

Semantic search emerged as a direct response to the limitations of lexical retrieval, shifting the objective from word matching to meaning matching. Early semantic approaches relied on manually constructed knowledge structures—ontologies, taxonomies, and concept graphs—to encode relationships between terms that keyword systems cannot infer [16]. While effective in narrow, well-defined domains, these approaches required extensive expert curation and did not scale gracefully to large, dynamic corpora.

The breakthrough came with distributional semantics, which holds that meaning can be inferred from patterns of co-occurrence across large text corpora. Word embedding models such as Word2Vec and GloVe demonstrated that words with similar meanings occupy proximate positions in a learned vector space. The introduction of the transformer architecture by [20] extended this principle dramatically. By incorporating attention mechanisms that model

relationships across entire input sequences, transformers enabled contextual embeddings: representations that encode not just a word's general meaning but its specific meaning in a given context.

BERT [4] applied this architecture to language understanding at scale, demonstrating that pretrained transformer models could be fine-tuned for a wide range of downstream tasks with state-of-the-art results. Research on the geometry of these embedding spaces has confirmed that contextual representations from BERT, ELMo, and GPT-2 capture semantic structure that static embeddings cannot, with higher layers encoding increasingly context-specific meanings [7]. [12] demonstrated that sentence-level embeddings derived from pretrained language models encode rich semantic information suitable for similarity-based retrieval, even without task-specific fine-tuning.

The application of these models to information retrieval crystallized with the development of dense passage retrieval [10], which demonstrated that transformer-based bi-encoder architectures substantially outperform BM25 on open-domain question answering benchmarks. This represented not merely an incremental improvement but a paradigm shift: retrieval systems could now be trained end-to-end to retrieve by meaning rather than by word.

2.3 Embedding Models and Vector Similarity Search

The practical implementation of semantic search requires not only embedding models capable of producing meaningful representations but also infrastructure capable of comparing them efficiently at scale. [17] addressed the former by introducing Sentence-BERT, a modification of BERT that produces sentence-level embeddings directly optimized for semantic similarity tasks. Sentence-BERT substantially outperforms averaging word-level BERT embeddings on standard semantic textual similarity benchmarks, making it the de facto foundation for practical semantic retrieval systems.

The latter challenge—efficient similarity search over large embedding collections—is addressed by approximate nearest neighbor (ANN) algorithms and vector database infrastructure. [8] demonstrated billion-scale similarity search using GPU-accelerated methods, establishing the computational feasibility of semantic retrieval at web scale. This infrastructure challenge remains one of the central engineering constraints distinguishing small-scale semantic search deployments from production systems: the mathematics of semantic retrieval scales far less gracefully than inverted index search, and this asymmetry has significant implications for educational platform adoption.

Interpretability represents an additional constraint that quantitative benchmarks do not capture. Keyword search produces rankings that are transparent: a retrieved result can be audited by examining which query terms matched which document fields. Semantic search rankings are based on vector similarity scores whose relationship to human judgments of relevance is not always intuitive [6]. This opacity has implications for educator and learner trust in retrieval systems, and for the ability of platform designers to diagnose and correct retrieval failures—a consideration that becomes particularly important in high-stakes educational contexts.

2.4 Semantic Search in Educational and Multimedia Contexts

The application of semantic retrieval to educational video content introduces challenges specific to the domain. Educational videos are multimodal artifacts: their meaning is carried simultaneously by spoken language, visual demonstrations, on-screen text, and structural cues such as pacing and segmentation. Most retrieval systems, including the system evaluated here, operate exclusively on textual representations—titles, descriptions, and transcripts—and therefore capture only a portion of the content's instructional meaning.

Transcripts present particular challenges. Automated speech recognition (ASR) systems, which generate transcripts for most large-scale educational platforms, introduce systematic error patterns: proper nouns and technical terms are frequently misrecognized, informal spoken phrasing differs structurally from written academic language, and disfluencies can disrupt sentence-level coherence. [13] demonstrated that embedding-based representations of transcripts and metadata can improve retrieval performance for complex queries, but noted that the extent of improvement varies with transcript quality. Semantic models are more robust to individual transcription errors than keyword systems—the attention mechanism can leverage surrounding context to recover meaning from corrupted input—but systematic misrecognition of domain-specific terminology can still degrade embedding quality.

Research specifically addressing educational video retrieval has followed two main tracks. The first, exemplified by [16], uses ontological and semantic indexing approaches to represent instructional content at the concept level. These methods demonstrate strong precision for structured educational domains but require manual knowledge engineering that limits scalability. The second track, represented by more recent neural approaches, uses data-driven embeddings to index video content at scale [18] - [5]. [5] demonstrated that dual-encoder architectures can retrieve semantically relevant video content even in zero-shot conditions—where queries and videos share no vocabulary overlap—substantially outperforming keyword baselines on this adversarial evaluation.

Despite these advances, direct controlled comparisons between semantic and keyword retrieval specifically for technical educational video search remain sparse in the literature. Most existing studies evaluate either retrieval methods in general video search contexts or semantic indexing methods without a keyword baseline for comparison. This study addresses that gap by implementing both approaches over the same dataset under controlled conditions and evaluating outcomes across query types directly relevant to the educational context.

2.5 Summary of Research Gaps

The literature establishes four key points. First, keyword search remains effective for terminology-precise queries but structurally disadvantages users who lack domain vocabulary—a failure mode that is especially consequential in educational settings. Second, transformer-based semantic retrieval has demonstrated consistent improvements over lexical methods for concept-driven and natural language queries across multiple domains [10] - [21]. Third, while semantic methods have been applied to educational and multimedia retrieval, direct controlled comparisons in the specific context of technical educational video search remain limited. Fourth, concerns about scalability, interpretability, and infrastructure cost [6] - [8] mean that retrieval quality alone cannot drive adoption decisions.

This study contributes to the literature by examining these dynamics in a domain—technical educational video retrieval—where the stakes of vocabulary mismatch are particularly high, where the gap between novice query language and expert indexed content is structurally widest, and where the equity implications of retrieval design have received insufficient attention. By evaluating both systems across three query categories under controlled conditions and situating the findings within the broader constraints of real-world deployment, this paper aims to advance both the empirical and normative conversation about how educational platforms should serve the learners they are designed for.

3. Methods and System Architecture

3.1 Study Design Overview

This study employs a controlled comparative evaluation design in which two retrieval systems—a keyword-based baseline and a semantic search system—operate over an identical dataset and respond to an identical query set, ensuring that observed differences in retrieval outcomes are attributable to the retrieval approach rather than variation in data or query conditions. The evaluation framework prioritizes conceptual alignment and learner intent over strict quantitative metrics, reflecting the centrality of instructional relevance in educational retrieval contexts [19].

This qualitative orientation is both a methodological choice and a recognized limitation. Qualitative evaluation of retrieval alignment enables nuanced assessment of whether retrieved content addresses the underlying learning goal of a query—a judgment that binary relevance metrics cannot fully capture—but it introduces subjectivity that limits reproducibility and generalizability. These limitations are addressed directly in Section 3.6 and Section 6. The study should be understood as a principled exploratory evaluation that generates hypotheses for future quantitative investigation rather than as a definitive empirical demonstration.

3.2 Dataset Compilation and Preprocessing

The dataset consists of publicly available educational videos focused on technical subjects including computer science, data structures, algorithms, engineering, and mathematics. Videos were drawn from MIT OpenCourseWare [15] and Khan Academy [11], spanning content from introductory programming concepts to intermediate-level treatments of algorithmic complexity and mathematical foundations of computing.The dataset encompasses videos of varying format and depth: short concept explainers (typically 5–15 minutes), extended lecture recordings (45–90 minutes), and structured tutorial series. This heterogeneity was intentional, reflecting the range of content a learner might encounter in a realistic search environment.

Transcripts were generated through automated speech recognition rather than manual transcription, consistent with the indexing approach used by major educational video platforms. This choice prioritizes ecological validity: a comparison using manually corrected transcripts would not reflect the conditions under which these systems would actually operate. Both the keyword and semantic systems were evaluated under identical transcript conditions, ensuring that any systematic transcription quality issues affect both approaches equally rather than creating an artificial advantage for either method. Prior research suggests semantic retrieval is

more robust to transcription noise—leveraging contextual embeddings to recover meaning from corrupted tokens—but the current design cannot isolate this effect [13].

Text preprocessing included lowercasing and removal of non-informative characters. Aggressive normalization such as stemming or stop-word removal was intentionally avoided to preserve contextual cues that contribute to semantic meaning, particularly for the embedding-based system.

3.3 Keyword-Based Search Baseline

The keyword-based retrieval system serves as the lower-bound baseline for the comparison. It indexes video titles, descriptions, and transcripts using a traditional inverted index and ranks results according to term frequency and inverse document frequency, approximating standard keyword retrieval mechanisms including TF–IDF and BM25 [14]. Queries are processed by tokenizing input text and retrieving documents that contain overlapping terms. No query expansion, synonym substitution, or semantic enrichment is applied, preserving the pure lexical matching behavior that characterizes standard keyword search.

This baseline was designed to represent the retrieval approach characteristic of most small-to-medium educational video platforms — such as Khan Academy, whose recent search updates have focused on keyword filters and subject domain categories rather than semantic retrieval — rather than the most effective possible keyword system. The absence of query expansion or pseudo-relevance feedback means the baseline may underperform relative to production keyword systems at major platforms, which is a recognized limitation of the comparison. However, it accurately represents the retrieval experience of users on platforms that have not invested in advanced keyword optimization.

3.4 Semantic Search System Architecture

The semantic search system converts all video text—titles, descriptions, and transcripts—into dense vector embeddings using a pretrained Sentence-BERT model [17]. Each video is represented by embeddings derived from its associated text fields, which serve as the system's semantic representation of the video's instructional content. Transformer models generate contextual embeddings that capture semantic relationships across entire text sequences, enabling the system to represent complex technical explanations at the concept level rather than the token level [21].

Embeddings are stored in a PostgreSQL database using the pgvector extension [1], which supports efficient vector similarity search within a relational database environment. Cosine similarity is used as the primary distance metric, as it performs well for comparing normalized text embeddings in retrieval tasks [17]. At query time, the query string is embedded using the same model and compared against stored video embeddings to retrieve the most semantically similar results. This architecture enables the system to identify conceptually relevant content even when queries and documents share no surface vocabulary.

Design Tradeoffs and Infrastructure Choices

PostgreSQL with pgvector was selected over dedicated vector databases such as Pinecone or Weaviate to prioritize simplicity, reproducibility, and integration with structured metadata.

Research on ad-hoc video search has shown that embedding-based retrieval can be made practically efficient within standard database infrastructure for moderate-sized corpora [22]. This choice reflects a deliberate design preference for transparency over maximum scalability—appropriate for an exploratory research evaluation but not necessarily appropriate for production deployment at scale. The scalability implications of this architectural choice are discussed in Section 5.

3.5 Query Design and Evaluation Framework

Thirty queries were constructed to reflect realistic learner information needs in technical educational search — a scale consistent with exploratory qualitative evaluation in retrieval research [19]. Queries were distributed across three categories: twelve natural language questions phrased as descriptive or explanatory queries; ten paraphrased conceptual queries expressing technical concepts using informal or non-standard vocabulary; and eight exact technical term queries using formal domain terminology. This distribution intentionally overweights the query types most likely to expose differences between keyword and semantic retrieval, while preserving a meaningful baseline category where keyword search is expected to perform well. An equal distribution across categories was considered but rejected on the grounds that the primary research question concerns retrieval behavior under vocabulary mismatch conditions.

Retrieval relevance was evaluated qualitatively by assessing whether the top-ranked result for each query directly addressed the conceptual intent of the query rather than merely sharing overlapping terms. Alignment with learner intent was determined by comparing the inferred learning goal of the query with the primary instructional focus of the retrieved video. For paraphrased queries, consistency was additionally evaluated by observing whether semantically equivalent queries in different linguistic forms returned conceptually equivalent results.

It is important to acknowledge two design limitations. First, queries across categories were not systematically constructed as parallel formulations of the same underlying concept. A more rigorous design would construct matched triplets—exact term, natural language, and paraphrased versions of the same information need—enabling direct within-concept comparison across query types. The absence of this parallel structure means category-level comparisons reflect differences in both retrieval approach and query construction, which are not fully separable in the current design. Second, the relatively short query lengths used in this study (2–8 words) may introduce bias favoring semantic search, as keyword systems typically perform better with longer, more specific queries. Future work should address both limitations.

Table 1 presents representative examples from the evaluation set illustrating observed differences in retrieval behavior across query categories. Table 2 presents illustrative frequency estimates summarizing the proportion of queries in each category for which the top result was judged intent-aligned. These estimates represent the researcher’s qualitative assessments and should be interpreted as characterizations of observed patterns rather than precise quantitative measurements.

Table 1

Qualitative Comparison of Retrieval Outcomes by Query Category

Query Type	Example Query	Keyword Result	Semantic Result	Observed Alignment
Natural Language	"Why does my code get slower when I add more loops?"	Video: loop syntax tutorial	Video: time complexity and algorithm efficiency	Semantic search shows substantially higher intent alignment
Natural Language	"What happens when a program runs out of memory?"	Video: memory allocation basics	Video: stack overflow, heap management, and memory errors	Semantic search retrieves conceptually precise content
Paraphrased Concept	"Making programs run faster"	General programming tips	Lecture: algorithm optimization techniques	Semantic search retrieves conceptually relevant content
Paraphrased Concept	"How recursion affects performance"	Video: recursion basics	Video: recursion and computational complexity	Semantic search maintains consistency across paraphrased queries
Exact Technical Term	"Big-O notation"	Big-O notation tutorial	Big-O notation tutorial	Both systems perform comparably
Exact Technical Term	"Binary search tree"	Binary search tree implementation	Binary search tree implementation	Both systems perform comparably

Note. Representative examples from the 30-query evaluation set. Results reflect the researcher's qualitative assessment of intent alignment. Queries were not systematically constructed as parallel formulations across categories.

Table 2

Illustrative Frequency Estimates of Intent-Aligned Top Results by Query Category

Query Category	Keyword: Intent-Aligned	Semantic: Intent-Aligned	Observed Difference	n (queries)
Natural Language	~40%	~85%	Semantic substantially higher	12
Paraphrased Concept	~35%	~80%	Semantic substantially higher	10
Exact Technical Term	~90%	~92%	Comparable performance	8
Overall	~53%	~85%	Semantic higher across set	30

Note. Percentages represent the researcher's qualitative assessments of the proportion of queries in each category for which the top-ranked result was judged intent-aligned. These are illustrative estimates, not precision metrics derived from annotated ground truth. They should be interpreted as characterizing observed patterns rather than as precise empirical measurements. n = number of queries per category.

3.6 Methodological Limitations

The evaluation design has three significant limitations that readers should bear in mind when interpreting the findings. First, all alignment judgments were made by a single researcher without a defined scoring rubric, inter-rater reliability check, or annotated gold standard. The subjectivity of these judgments means that a different evaluator might reach different conclusions about the same retrieval outcomes. The frequency estimates in Table 2 in particular should be understood as characterizations of the researcher's qualitative impressions rather than reproducible quantitative measurements.

Second, the study lacks an upper-bound benchmark. By comparing semantic search against a keyword baseline, the evaluation establishes that semantic search performs better than keyword search for certain query types, but it does not establish how close either system comes to optimal retrieval. Establishing an upper bound would require expert educators to annotate relevance judgments for each query-video pair, which was beyond the scope of this evaluation.

Third, as noted in Section 3.5, the non-parallel query design and short query lengths may confound query category effects with retrieval approach effects. Future studies should address this by constructing matched parallel query sets and evaluating with longer, more naturalistic learner-generated queries.

4. Results

4.1 Overview

Across the thirty-query evaluation set, the semantic search system demonstrated substantially higher intent alignment than the keyword baseline for natural language and paraphrased conceptual queries, while both systems performed comparably for exact technical term queries. Table 2 summarizes these patterns: semantic search produced intent-aligned top results for approximately 85% of natural language queries and 80% of paraphrased queries, compared to approximately 40% and 35% for keyword search respectively. For exact technical term queries, the two systems were nearly indistinguishable (~92% vs. ~90%).

These patterns are consistent across individual query examples in Table 1 and across the broader evaluation set from which those examples were drawn. The discussion that follows characterizes the qualitative differences in retrieval behavior observed across each category. All characterizations reflect the researcher's interpretive judgments based on observed retrieval behavior; they should be weighed in light of the methodological limitations described in Section 3.6.

4.2 Performance on Exact Technical Term Queries

When queries contained precise, formally defined technical terminology—"Big-O notation," "binary search tree," "Dijkstra's algorithm"—both systems consistently retrieved intent-aligned results. The keyword system succeeded because strong lexical overlap between query terms and indexed content provided reliable retrieval signals. The semantic system produced equivalent results because contextual embeddings of formal technical terms are well-defined and cluster tightly in the embedding space around related instructional content.

The near-equivalence of the two systems in this category is itself informative. It confirms that semantic search does not introduce degradation for the query type where keyword search is strongest, meaning the case for semantic search rests on its advantages in other categories rather than requiring trade-offs in this one. This supports the broader argument for hybrid retrieval architectures that leverage semantic search where vocabulary mismatch is most acute without sacrificing keyword search efficiency where lexical precision is adequate.

4.3 Performance on Natural Language Queries

The sharpest divergence between systems emerged in the natural language query category. Queries phrased as descriptive or explanatory questions—"Why does my code get slower when I add more loops?" or "What happens when a program runs out of memory?"—represent the kind of search behavior that characterizes novice learners navigating unfamiliar conceptual terrain. These queries describe a phenomenon or express a confusion; they do not name the concept the learner is seeking.

The keyword system retrieved content that shared surface vocabulary with the query but frequently missed the underlying concept. A query about code getting slower with more loops returned videos about loop syntax; a query about programs running out of memory returned videos about memory allocation basics. In both cases, the retrieved content was technically

related to the query terms but did not address the learner's actual information need. The system matched words while missing meaning.

The semantic system, by contrast, consistently identified the conceptual target behind the informal query language. The time complexity and algorithm efficiency video—the conceptually correct result for the loop-slowness query—does not necessarily contain the phrase "gets slower" but is embedded in a region of the semantic space that corresponds to computational performance and algorithmic analysis. The memory error video similarly retrieved content addressing stack overflows and heap exhaustion—the actual concepts behind the learner's question—rather than introductory memory allocation tutorials. Across the natural language category, this pattern held consistently: semantic search bridged the gap between how learners ask and what they need.

4.4 Performance on Paraphrased Conceptual Queries

Paraphrased conceptual queries test a different dimension of retrieval quality: consistency across linguistic variation. Two queries that express the same underlying concept using different vocabulary should ideally retrieve the same instructional content. Keyword search fundamentally cannot guarantee this, because it treats different word choices as different retrieval signals. Semantic search can, because it encodes conceptual similarity rather than lexical identity.

The evaluation confirmed this theoretical difference empirically. Keyword search produced inconsistent results across paraphrased query pairs, often retrieving fundamentally different content for queries that targeted the same concept. "Making programs run faster" and "how recursion affects performance"—both pointing toward algorithmic efficiency and computational complexity—returned unrelated results under keyword search. The semantic system maintained coherent retrieval across both queries, returning content in the algorithmic complexity and optimization domain regardless of how the concept was expressed.

This consistency has direct implications for educational platform design. If a learner's first phrasing of a query returns unhelpful results and they rephrase it, they should not receive an entirely different and equally unhelpful set of results. Keyword search offers no guarantee of this; semantic search, by encoding meaning rather than tokens, substantially improves retrieval consistency across the natural variation in how learners express the same information need.

4.5 Comparative Summary

The results present a coherent and consistent pattern: semantic search substantially outperforms keyword search precisely where keyword search is most structurally limited—informal, concept-driven, naturally expressed queries—while performing equivalently where keyword search is strongest—exact technical terminology. This pattern is not surprising given the theoretical underpinnings of each approach, but it is significant: it confirms that the theoretical advantages of semantic retrieval manifest in practice within the specific domain of technical educational video search, and it clarifies the conditions under which each approach is appropriate.

The equity implications of this pattern deserve emphasis. The query types for which semantic search most outperforms keyword search—natural language and paraphrased conceptual queries—are precisely the query types most characteristic of novice learners. Expert learners

who command technical vocabulary can formulate effective keyword queries; they are the users for whom keyword search already works. Novice learners, who need instructional support most urgently, are the users keyword search most reliably fails. Semantic search inverts this dynamic, providing better retrieval exactly where better retrieval is most consequential.

5. Discussion

The central finding of this study—that semantic search substantially outperforms keyword search for the query types most characteristic of novice learners—reframes a technical comparison as an equity question. The choice of retrieval architecture in an educational platform is a choice about which learners the platform effectively serves. A platform that indexes rich instructional content but retrieves it only for users who already know the right words is not democratizing education; it is replicating existing knowledge barriers in digital form.

This framing builds on and extends the vocabulary mismatch literature [19] - [3] by making explicit the distributional consequences of keyword-based retrieval design. Vocabulary mismatch is not a random failure mode that affects all learners equally—it disproportionately affects those at the frontier of their conceptual development, users who are reaching for understanding they do not yet have and who cannot name what they are looking for because the naming is precisely what they are trying to learn. Semantic search addresses this not by being a better search engine in the abstract, but by being a better interface for the specific cognitive situation of the learner.

The practical implications for platform design are direct. For platforms serving primarily advanced learners or expert practitioners, keyword search remains an efficient and effective default, with semantic search adding value only at the margins. For platforms serving novice learners navigating technical material—introductory CS courses, self-directed technical learners, adult learners developing new skills—the case for semantic search is substantially stronger. The strongest design recommendation from this study is the hybrid architecture: use keyword search for exact-term queries where lexical precision provides adequate signal, and route natural language and ambiguous queries through the semantic system where meaning inference is required [21].

That said, the path from principle to production is not straightforward. Deploying semantic search at the scale of major educational platforms introduces infrastructure challenges that cannot be dismissed. Generating and storing dense embedding vectors requires GPU computation at index time and efficient vector search infrastructure at query time [8]. For platforms hosting hundreds of thousands of videos and serving millions of concurrent users, the computational cost of semantic indexing and retrieval—particularly for real-time query processing—is substantially higher than for inverted index search, which scales linearly with corpus size and serves queries in microseconds. [10] document the engineering complexity of dense retrieval at scale, noting that production deployment requires careful optimization of both embedding generation pipelines and ANN search parameters. This cost-benefit asymmetry explains why keyword-based systems remain dominant in production despite the well-documented retrieval quality advantages of semantic approaches.

Interpretability presents a second practical constraint. Keyword search is auditable: when a retrieved result seems wrong, a developer or educator can inspect which query terms matched which document fields and diagnose the failure. Semantic search rankings are grounded in cosine distances between high-dimensional vectors—mathematically well-defined but not intuitively interpretable [6]. When a semantic system retrieves a video that seems wrong, understanding why requires either examining the embedding space directly (a technical operation inaccessible to most educators and learners) or accepting the system's judgment on faith. As educational platforms become more algorithmically mediated, the ability of educators, learners, and administrators to understand and contest retrieval decisions becomes increasingly important. Addressing this through interpretable semantic retrieval—explaining rankings in terms of matched concepts rather than vector distances—is an important research direction that this study cannot resolve but that future work should prioritize.

Finally, the evaluation methodology used in this study requires honest reflection. Qualitative alignment judgment by a single researcher, while appropriate for exploratory evaluation and sensitive to instructional value in ways that precision metrics are not, does not produce evidence strong enough to drive platform-level design decisions. The frequency estimates in Table 2 should be understood as characterizations of qualitative impressions rather than rigorous measurements. The conclusion that semantic search substantially outperforms keyword search for concept-driven queries is supported by the pattern of evidence—but it is a hypothesis strengthened by this study, not a theorem proven by it. Confirming it at the level of rigor required for policy recommendations will require annotated relevance datasets, controlled user studies, and learner outcome measurements that this work does not provide.

6. Limitations and Proposed Solutions

This study has five significant limitations that collectively bound the confidence with which its conclusions can be generalized.

First and most fundamentally, all relevance judgments were made by a single researcher using qualitative assessment without defined scoring rubrics or inter-rater reliability measures. The frequency estimates in Table 2 reflect the researcher's impressions rather than reproducible measurements. Different evaluators with different domain expertise or pedagogical values might assess alignment differently, particularly for borderline cases where a retrieved video is partially relevant to a query. Future research should develop multi-annotator evaluation protocols with explicit relevance criteria calibrated to educational value, and should report inter-annotator agreement to establish the reliability of alignment judgments.

Second, the study lacks a gold standard that would establish what optimal retrieval looks like for this query set. The evaluation demonstrates that semantic search outperforms the keyword baseline, but it does not establish how far either system is from optimal. Creating such a benchmark requires subject-matter experts to annotate query-video relevance for the full evaluation set, accounting for instructional alignment, conceptual accuracy, and learning progression—a task that demands both technical and pedagogical expertise.

Third, the query design has two structural weaknesses: queries across categories were not constructed as parallel formulations of the same information need, and query lengths (2–8 words) may favor semantic search relative to longer, more specific queries where keyword systems perform better. Future studies should construct matched query triplets—exact term, natural language, and paraphrased versions of the same concept—and should include learner-generated queries collected in naturalistic search contexts rather than researcher-constructed queries.

Fourth, the comparison does not test the most sophisticated possible keyword system. A keyword baseline augmented with query expansion, pseudo-relevance feedback, or synonym substitution might close some of the performance gap observed here. The keyword baseline used in this study represents what most small-to-medium educational platforms provide in practice, not the upper bound of what lexical retrieval can achieve.

Fifth, the scalability and interpretability constraints discussed in Section 5 remain unresolved. The study demonstrates the retrieval quality case for semantic search but does not address how to realize those benefits at production scale or how to make semantic rankings explainable to the educators and learners who depend on them.

These limitations point toward a concrete research agenda: develop expert-annotated evaluation datasets for technical educational video retrieval; conduct controlled user studies that measure both retrieval satisfaction and learning outcomes; investigate hybrid architectures that combine keyword and semantic retrieval; and explore techniques for making semantic retrieval more interpretable. Until these gaps are addressed, the conclusions of this study should be understood as well-motivated hypotheses rather than established findings.

7. Conclusion

This study evaluated semantic search against keyword-based retrieval for educational video discovery in technical subjects. Across thirty queries spanning three categories, semantic search consistently produced more intent-aligned results for natural language and paraphrased conceptual queries—the query types most characteristic of novice learners—while both systems performed equivalently for exact technical terms. These patterns suggest that the vocabulary mismatch problem documented in the information retrieval literature manifests in technical educational video search in a form with direct equity implications: keyword-based retrieval systematically advantages learners who already possess domain vocabulary, while semantic retrieval extends effective search to learners who are still developing it.

The central contribution of this paper is not technical but conceptual: it reframes retrieval architecture as a pedagogical design decision with distributional consequences. The question of whether a platform deploys keyword or semantic search is a question about which learners it effectively serves. Platforms that adopt keyword search by default—as most do—are not making a neutral technical choice. They are making a choice that works well for expert users and poorly for novices, which is the opposite of what a platform explicitly designed for learning should optimize for.

At the same time, this paper has been honest about the limits of what its evidence can support. Qualitative single-researcher evaluation is appropriate for exploratory investigation but cannot bear the evidential weight of policy recommendations. Realizing the equity benefits of semantic search in practice requires addressing the genuine infrastructure, interpretability, and evaluation challenges that this study surfaces but cannot resolve. The path from the promising patterns documented here to production deployment at platforms like Coursera or Khan Academy runs through controlled user studies, annotated relevance benchmarks, domain-adapted embedding models, and architectural innovation in hybrid retrieval.

The ultimate measure of a retrieval system in an educational context is not recall or precision on a benchmark dataset. It is whether learners who do not know what to search for can find what they need to learn. By that measure, this study suggests semantic search has a meaningful advantage—and that closing the gap between that advantage in principle and its realization in practice is work worth doing.

References

[1] Ankeny, A. (2024). pgvector (Version 0.8.0) [PostgreSQL extension]. GitHub. https://github.com/pgvector/pgvector

[2] Çakir, H., Acartürk, C., Alaşehir, O., & Çilingir, C. (2018). Improving educational web search for question-like queries through subject classification. Information Processing & Management, 54(6), 1123–1138. https://doi.org/10.1016/j.ipm.2018.06.005

[3] Chawan, P. M., & Malve, A. (2015). A comparative study of keyword-based and semantic-based search engines. International Journal of Innovative Research in Science, Engineering and Technology, 4(6), 4219–4225. https://www.researchgate.net/publication/316514673

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[5] Dong, J., Li, X., Xu, Y., Ji, S., He, Y., & Yang, Y. (2018). Dual encoding for zero-example video retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 9346–9355). IEEE. https://arxiv.org/abs/1809.06181

[6] Eller, D. W. (2022). Transparency and the future of semantic searching. Information Services & Use, 42(4), 389–401. https://doi.org/10.3233/ISU-220175

[7] Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint

Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 55–65). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1006

[8] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

[9] Jurafsky, D., & Martin, J. H. (2023). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed. draft). Stanford University. https://web.stanford.edu/~jurafsky/slp3/

[10] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

[11] Khan, S. (2024). *Khan Academy* [Video platform]. https://www.khanacademy.org

[12] Li, B., Zhou, H., He, J., Wang, M., Yang, Y., & Li, L. (2020). On the sentence embeddings from pre-trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 9119–9130). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.733

[13] Lin, D., Fidler, S., Kong, C., & Urtasun, R. (2014). Visual semantic search: Retrieving videos via complex textual queries. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1–8). IEEE. https://www.cs.toronto.edu/~fidler/papers/lin_et_al_cvpr14.pdf

[14] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. https://nlp.stanford.edu/IR-book/

[15] Massachusetts Institute of Technology. (2024). *MIT OpenCourseWare* [Open educational resource]. https://ocw.mit.edu

[16] Merzougui, G., Djoudi, M., & Behaz, A. (2012). Conception and use of ontologies for indexing and searching by semantic contents of video courses. arXiv preprint. https://arxiv.org/abs/1201.5102

[17] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

[18] Stoica, A. S., Barbu, T., & Breaban, M. (2021). Classification of educational videos using a semi-supervised method. Neural Networks, 144, 487–498. https://doi.org/10.1016/j.neunet.2021.09.019

[19] Toriah, S., Ghalwash, A., & Youssif, A. (2018). Semantic-based video retrieval: A survey. Journal of Computer and Communications, 6(8), 1–15. https://www.scirp.org/journal/paperinformation.aspx?paperid=86488

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).

[21] Veluru, S. R., Marella, V. C., & Erukude, S. T. (2025). The evolution of search engines: From keyword matching to AI-powered understanding. SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5403467

[22] Wu, Y., & Ngo, C. W. (2024). Interpretable embedding for ad-hoc video search. arXiv preprint. https://arxiv.org/abs/2402.11812