MIXPEEK LEARNING INTELLIGENCE

# Turn Courses Into Machine-Readable Intelligence

Search, audit, and update curriculum at scale. Built for Learning Engineering teams.

| | | | |
|---|---|---|---|
| 79.3% | <200ms | 60% | 12x |
| NDCG@10<br>Retrieval Quality | P95 Latency<br>Real-time Search | Reduction<br>SME Costs | Faster Updates<br>6-12mo → 2wk |

## The Learning Engineering Crisis

Every learning platform faces the same challenges: **Content velocity** outpaces manual audits. **Catalog sprawl** makes search useless. **SME bottlenecks** create 6-12 month update cycles. Students expect ChatGPT-grade search while you deliver 2015-era keyword matching.

| | | |
|---|---|---|
| 40-60% | 6-12 months | $500K+ |
| of catalog content<br>is stale or outdated | lag between breaking<br>changes and updates | annual SME cost<br>for manual maintenance |

## The Solution: A Learning Intelligence Layer

Infrastructure that treats curriculum as **structured, queryable, version-controlled data**. Not files. Not videos. **Intelligence.**

### Multi-Modal Processing

- Video: Whisper ASR + Scene Detection
- Slides: PDF Processing + OCR
- Code: Multi-language Analysis + AST

### State-of-the-Art Retrieval

- Multi-Vector Embeddings (BGE-M3)
- HyDE Query Enhancement
- Reciprocal Rank Fusion

## SOTA Benchmark Results

Tested on CS50 curriculum (Harvard). Validated against gold-standard evaluation methodology.

| Metric | Mixpeek | Vector Only | BM25 | Target |
|---|---|---|---|---|
| NDCG@10 | **79.3%** | 68.2% | 54.7% | >75% |
| Recall@50 | **80%** | 65% | 52% | >90% |
| Latency p95 | **<200ms** | ~50ms | ~30ms | <200ms |

## What You Can Build

### Content Freshness Engine

Automatically detect when libraries, APIs, or vendor docs change. Flag outdated lecture segments without manual audits. Surface exact timestamps for SME review.

**→ 60% reduction in maintenance cost**

### Lecture Segment Search

Enable semantic search across millions of lecture minutes. Students find exact answers in seconds instead of opening support tickets asking "where is X explained?"

**→ 40% reduction in support load**

### AI Tutor Grounding Layer

Power LLM-based tutors with retrieval grounded in your actual curriculum. No hallucinations—every answer cites exact lecture moments and slide numbers.

**→ Trustworthy course chatbots**

### Taxonomy Alignment

Generate topic maps, skill tags, and learning objective metadata automatically. Build a curriculum graph showing which lectures cover "async/await" or "gradient descent."

**→ Skills mapped at scale**

## Built For

**Learning Engineering Teams** at Coursera, LinkedIn Learning, Pluralsight • **Content Tech / Platform Teams** at O'Reilly, Udacity, Khan Academy • **Certification Programs** tracking AWS, Azure, GCP changes • **Enterprise L&D;** centralizing Loom, Zoom, onboarding decks

## How It Works

1. **INGEST** → Video lectures, slide decks, code examples, documentation
2. **EXTRACT** → Whisper ASR • Scene detection • OCR • Code analysis
3. **EMBED** → Multi-vector: transcript, code, visual, bound context
4. **INDEX** → Vector store (Qdrant-ready) with metadata filtering
5. **RETRIEVE** → HyDE enhancement → Multi-vector search → RRF
6. **SERVE** → <200ms responses with timestamps, slides, code

### Example: Semantic Search in Action

**Query:** "What is memory allocation in C?"

**[1] Score: 0.8234** • Scene: 120.5-185.3s • Course: CS50 Lecture 4
"Memory allocation in C allows you to dynamically request memory from the heap using malloc..."

**Code:** `char *s = malloc(4); strcpy(s, "hi!");`

## Why Standard RAG Falls Short

Educational content requires specialized processing that general-purpose RAG can't deliver:

- ✗ **Temporal Misalignment**
  LLMs can't understand *when* concepts appear in 90-minute lectures
- ✗ **Multi-Modal Context Loss**
  Code on slides + audio explanation gets lost in single-vector search
- ✗ **Code Semantic Blindness**
  General embeddings don't understand imports, APIs, or structure
- ✗ **Vague Query Handling**
  "memory stuff" doesn't match exact transcript text without HyDE

### Mixpeek Solution

- ✓ Scene detection with word-level timestamps ensures precise temporal alignment
- ✓ Multi-vector representation maintains separate embeddings for each modality
- ✓ Specialized code embeddings (StarCoder/SFR) with AST analysis
- ✓ HyDE generates hypothetical explanations to improve query-content matching

| 15-20% | <200ms | 95%+ |
|---|---|---|
| Better retrieval vs single-vector | Multi-vector fusion latency | Scene-transcript alignment |

## Start a Learning Intelligence Pilot

**2 weeks. One slice of your catalog. See what's possible.**

No vendor lock-in. Open benchmarks. Production-ready infrastructure.

Trusted by Learning Engineering teams managing millions of lecture minutes at Coursera, LinkedIn Learning, and enterprise academies.

**mxp.co/learning**

ethan@mixpeek.com

• 915 Broadway, Suite 1200, New York, NY 10010 • mixpeek.com