Topic-Aware Semantic Segmentation and Learning Objective Induction

from Educational Content using Large Language Models

Thomas Plangger, BSc

Master's Thesis Defence · Graz University of Technology · March 2026

Supervisors: Priv.-Doz. Dr. Martin Ebner · Dr. Benedikt Brünner · Dr. Philipp Leitner

Presentation Outline

01 **Motivation & Problem**
Why structure matters in educational PDFs

02 **Research Questions**
What the thesis sets out to answer

03 **System Architecture**
End-to-end pipeline overview

04 **Semantic Chunking**
LLM-driven boundary detection

05 **LO Induction**
Grouping chunks into Learning Objectives

06 **Evaluation Setup**
Synthetic benchmark & experiments

07 **Results**
Chunking & LO grouping findings

08 **Conclusion**
Key takeaways and future work

Motivation

How can we make learning faster, more efficient, and more accessible for students and instructors?

Problem

Learning materials are often distributed as PDFs or slides, formats that work well for delivery but offer limited interactivity for students.

Personal Experience

During my studies, I found that using AI tools to get explanations, test my knowledge, and understand my mistakes made learning feel more engaging.

Automation

Build a hybrid system that generates quizzes for students **and** lets instructors create structured courses from their documents.

Goal: An end-to-end pipeline that takes educational PDFs and automatically produces structured courses with learning objectives, lessons, and quizzes making learning active and course creation scalable.

Why Focus on Chunking & LO Induction?

The most important part of the pipeline is deciding where topics begin and end and how they relate.

Semantic Chunking

- Reliable segmentation is essential because all later stages depend on correctly identifying topical units.
- Heading-based splitting is unreliable as documents might not have no headings.
- Rule-based approaches often struggle on heterogeneous content.
- PDF texts often have no machine-readable topic markers.

LO Induction

- Grouping related chunks into coherent learning objectives is necessary for structuring instructional material.
- Grouping chunks into pedagogically meaningful units requires understanding of content.
- Heuristic clustering (k-means) ignores meaning, only distances in embedding space.
- Minimal reproducible benchmark existed for either task, evaluation was the open gap.

Research Questions

RQ 1

Chunking Behaviour

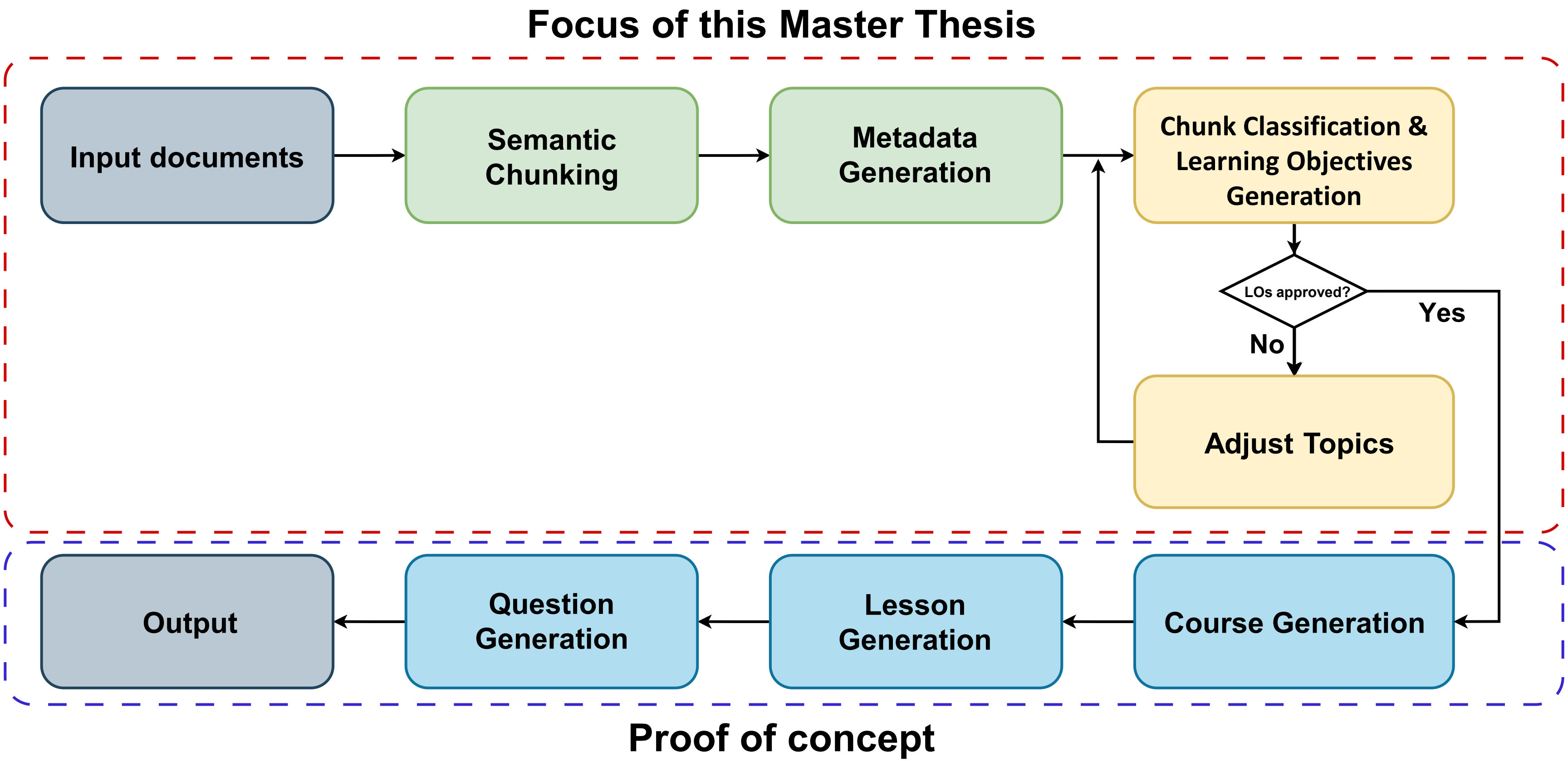
How do LLM-based chunking strategies behave under controlled, reproducible evaluation conditions, and to what extent do they produce boundaries that align with meaningful structural transitions in text?

RQ 2

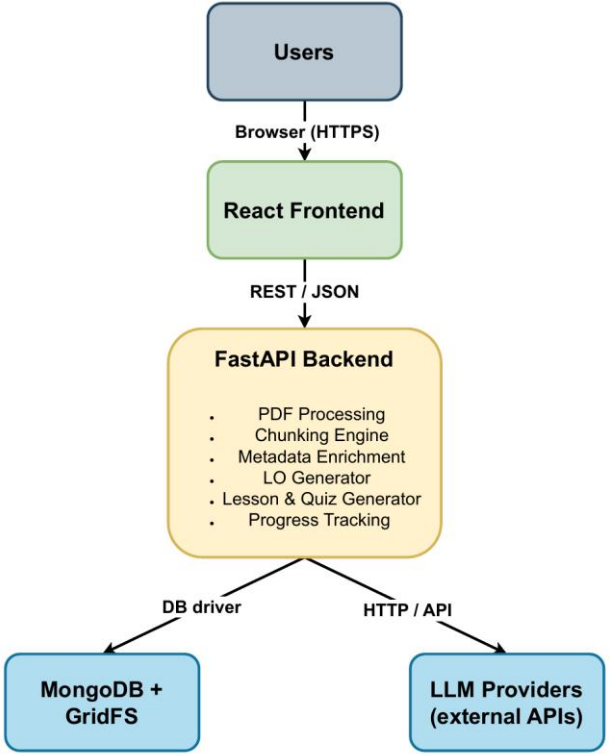
LO Quality & Structure

How well do automatically generated Learning Objectives organise chunked content into coherent higher-level units, and how does segmentation influence the structure of the resulting learning materials?

System Pipeline Overview



System Architecture



React Frontend

Browser-based UI · REST/JSON communication

FastAPI Backend

PDF processing · Chunking · LO · Lesson & Quiz generator

MongoDB + GridFS

Persistent storage of docs, chunks & artefacts

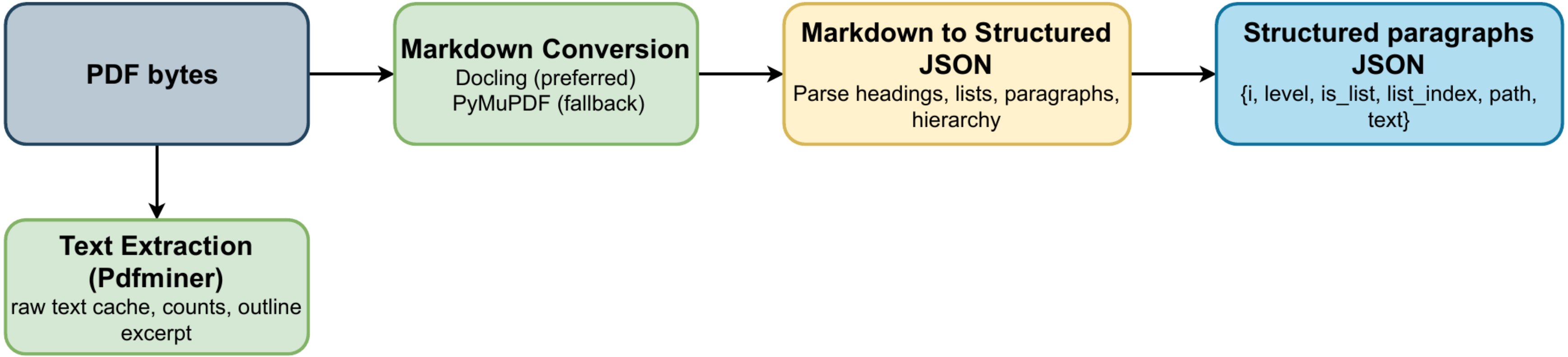
LLM Providers

OpenAI GPT-4 via external API · prompt-based inference

sentence-transformers

Embeddings (all-MiniLM-L6-v2) for semantic similarity

Step 1 — Input & Pre-processing



Text Extraction (pdfminer)

Raw text, character counts, outline excerpt, used for heuristic pre-checks and metadata enrichment.

Markdown Conversion (Docling)

Preferred converter: preserves heading hierarchy, lists, table structure. PyMuPDF used as fallback.

Structured JSON

Each paragraph encoded in the JSON file and the input is fed to the chunker.

Step 1 — Chunk Data Structure

```
[
  { "i": 0,  "level": "H3",  "is_list": false, "list_index": null,
    "path": [],  "text": "Mechanical Keyboard Tuning & Maintenance …" },  ← root heading
  { "i": 1,  "level": "P",  "is_list": false,  "list_index": null,
    "path": [0],  "text": "A structured, step-by-step handbook …" },  ← body paragraph
  { "i": 3,  "level": "H3",  "is_list": true, , "list_index": 0,
    "path": [0, 2],  "text": "1) Layout & Form Factor …" },  ← nested sub-section
  … (N more entries)
]
```

`i` Chunk index

`level` H1–H6 or P

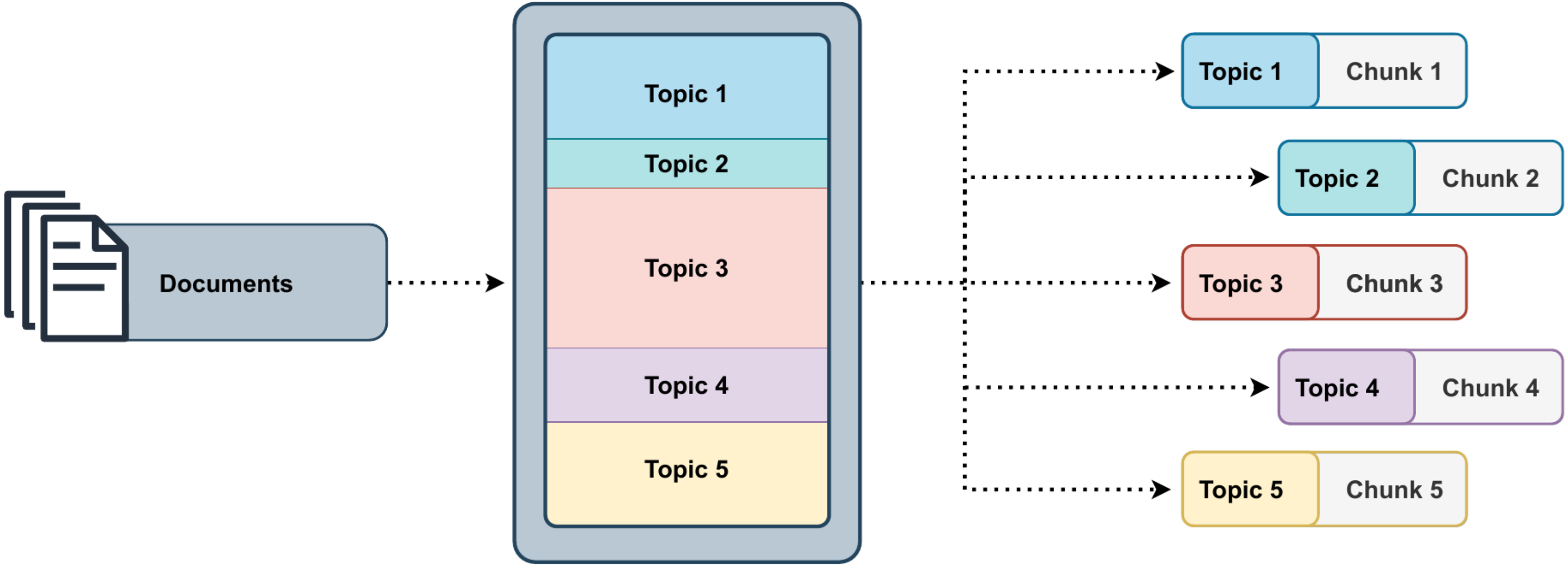
`is_list` List item flag

`path` Text hierarchy

`list_index` List index

`text` Raw paragraph text

Step 2 — Semantic Chunking



Hybrid approach: Heuristic pre-chunking + LLM boundary prediction (GPT-4) + post-processing rules (merge / split / size constraints)

Step 2 — Chunking Pipeline Detail

JSON Chunk File
Parsed paragraphs with index, level, path, text

Chunking Prompt
Instructions for grouping paragraphs into semantic chunks

Provider: ChatGPT API (AI Chunker)
Groups paragraph indices into semantic chunks; optionally generates metadata (title, summary, keywords)

Chunk Indices + Metadata
List of index values
title, summary, keywords per chunk

Index → Text Mapping
Each index maps back to the raw paragraph text in the source

Stored Chunks
Raw text chunks stored for reference
downstream LO induction uses indices + metadata primarily

Key Points
- Input is the flat JSON paragraph list from Step 2a
- JSON + prompt sent to ChatGPT API; returns grouped indices
- Optionally generates metadata: title, summary, keywords per chunk
- Index mapping resolves each "i" back to its source text location
- Raw text chunks stored but not primary; indices + metadata drive LO induction

Step 3 — Learning Objective Induction

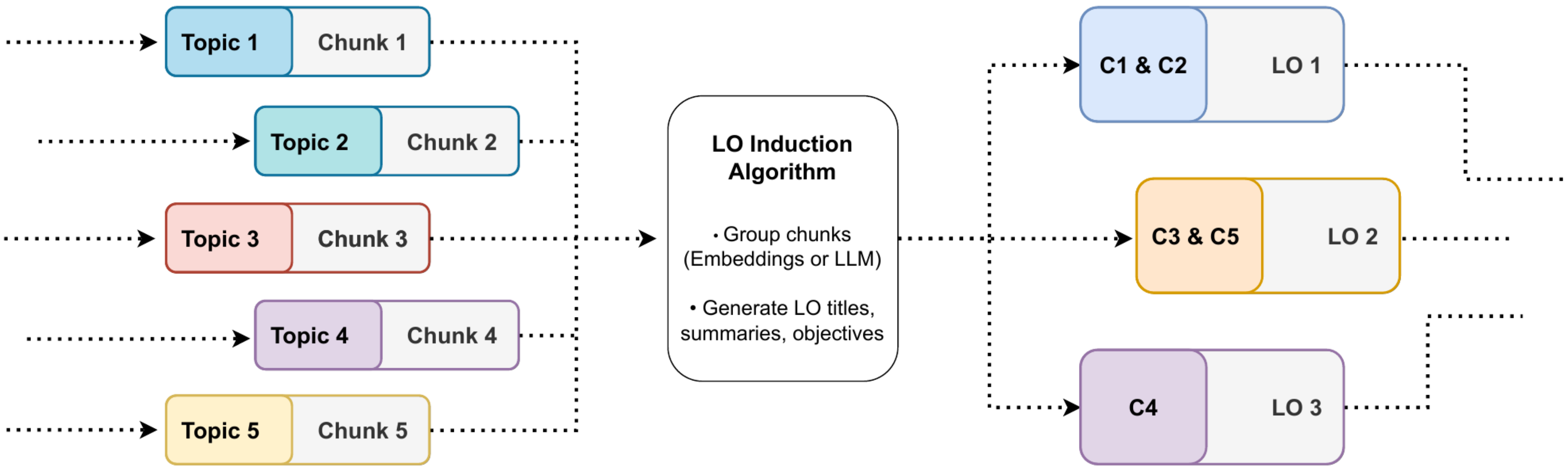


Fig: Chunks grouped into coherent Learning Objectives via the LO Induction Algorithm

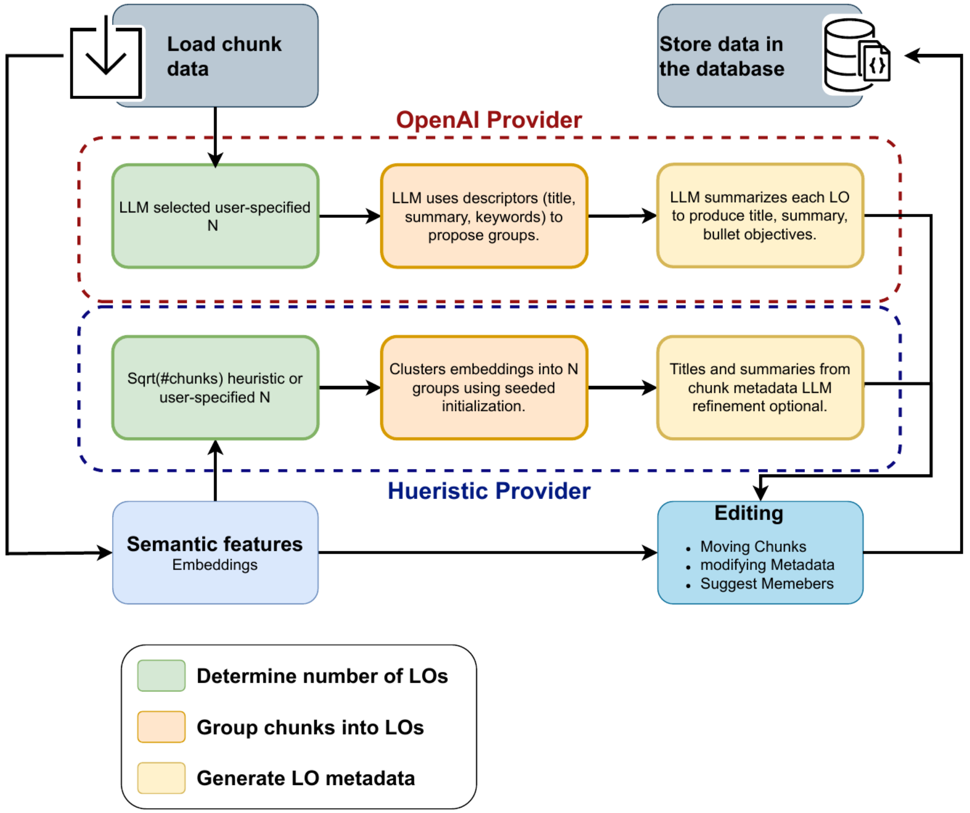
LLM-Based (OpenAI)

Chunk descriptors (title, summary, keywords) fed to GPT-4 → proposes semantic groups → generates LO title, summary, bullet objectives per group.

Heuristic (k-means)

√(#chunks) groups via seeded k-means on sentence embeddings → titles and summaries from chunk metadata; LLM refinement optional.

Step 3 — LO Pipeline Detail



Semantic Grouping

Chunk descriptors fed to provider (OpenAI or k-means) for grouping.

Metadata Generation

LO title, summary, and objectives auto-generated per group.

User Editing

Move chunks, edit metadata, re-suggest members via UI.

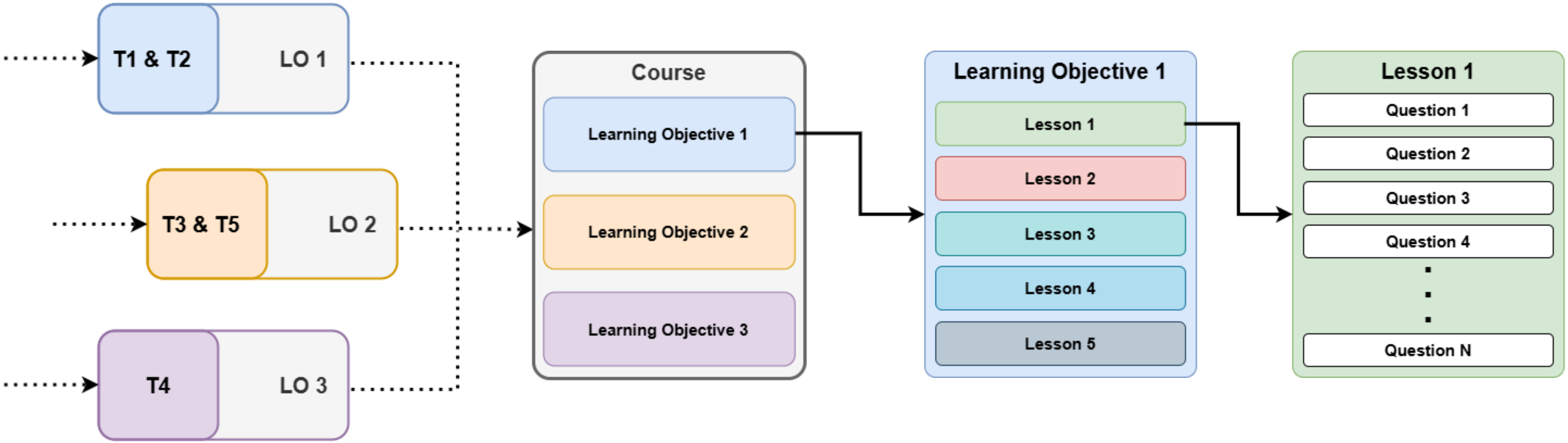
Dual Provider Design

Enables direct comparison of LLM vs. heuristic approaches.

Pipeline Output

Structured LOs ready for course generation in Step 4.

Step 4 — Course, Lesson & Question Generation (Proof of Concept)



Course Generation

A course container is built from the set of approved LOs. Each LO becomes a module.

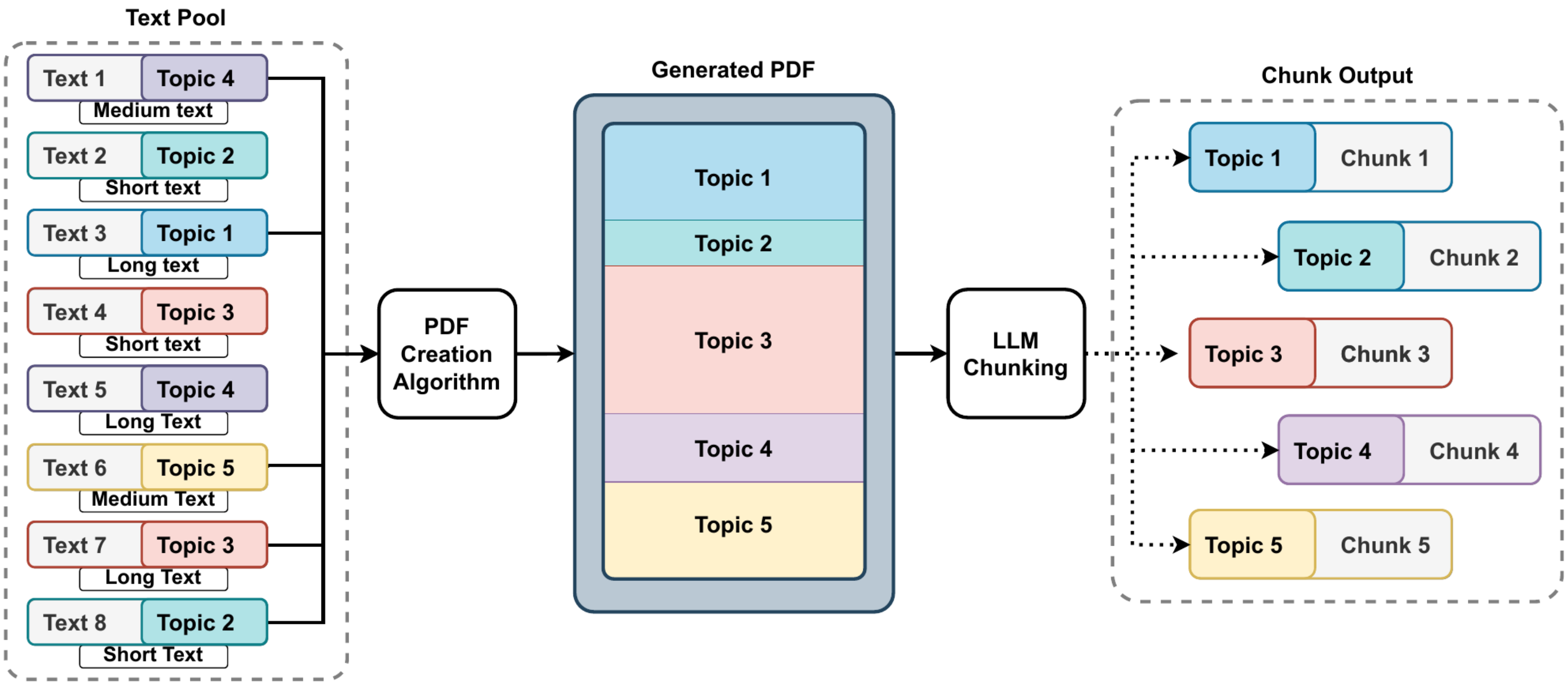
Lesson Generation

LOs get split into uniformly sized lessons. LLM writes title, summary, 2–3 paragraphs and key points per lesson.

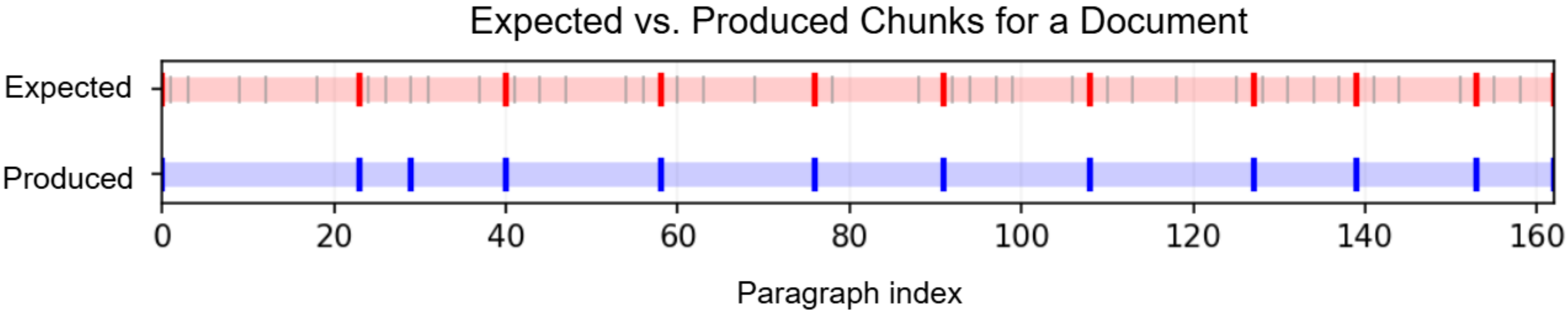
Question Generation

Per lesson: MCQ + free-text questions. LLM grades free-text answers against reference content.

Evaluation Setup — Auto-test Pipeline



Evaluation Framework — Boundary Visualisation



Synthetic Document Generation

LLM generates multi-topic test documents with controlled structure, length, and topic count

Gold-Standard Metadata

Metadata file stores true section boundaries, topic labels, and paragraph positions as ground truth

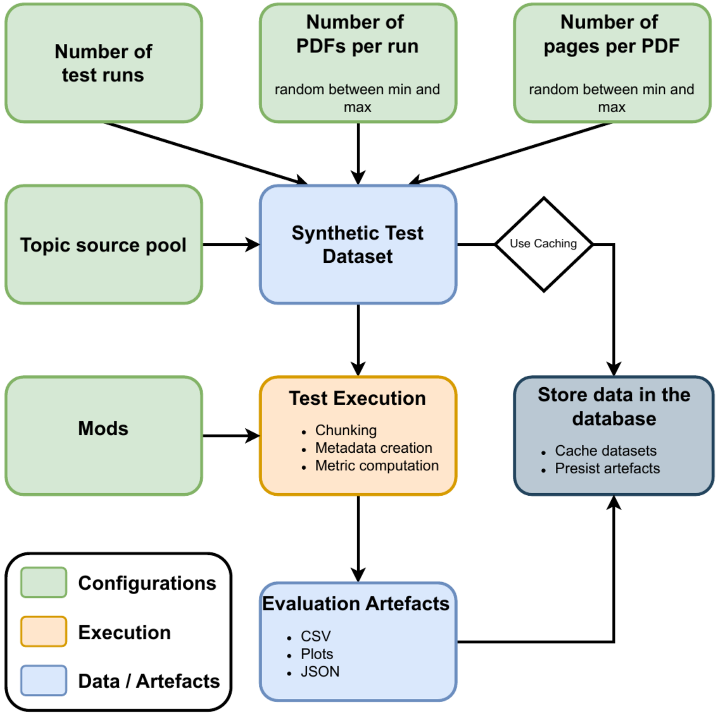
Chunking & Comparison

Predicted chunk boundaries are aligned against gold boundaries to compute precision, recall, and F1

Visual Inspection

Plot overlays expected vs. produced boundaries visualizing

Evaluation Setup — Synthetic Benchmark



Experiments A–F

A Doc Length
Scaling behaviour as documents grow

B Chunk Count
Over- vs under-segmentation detection

C Section Variability
Mixed short and long sections

D Topic Similarity
Homogeneous vs heterogeneous topics

E Formatting Noise
Missing headings, irregular structure

F Post-Processing
Effect of merge/split refinement rules

Key Metrics

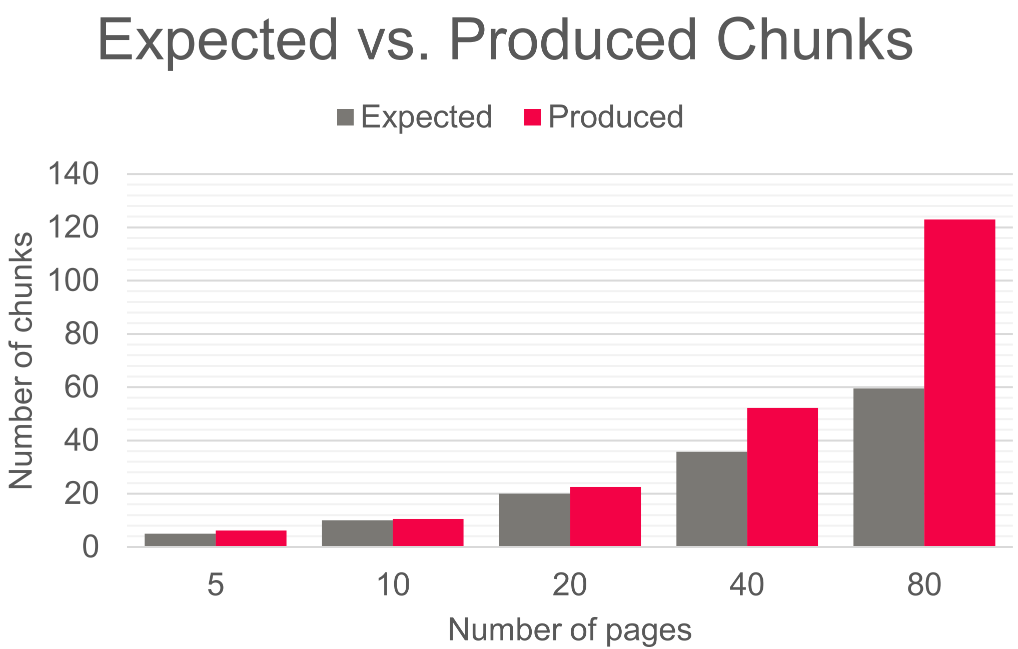
Precision / Recall
Boundary correctness & completeness

F1 Score
Balanced segmentation quality (main metric)

Over/Under-Seg
Too many or too few chunks detected

Synthetic PDFs with known boundaries → reproducible, quantitative evaluation across 6 experiment types (A–F)

Results — Chunking Behaviour



Sample size: 125 Test runs

Over-segmentation

F1 drops with document length, over-segmentation is the biggest problem at longer documents

LLM vs. Heuristic

LLM-based chunking outperforms heuristic baseline based on the TextTiling algorithm

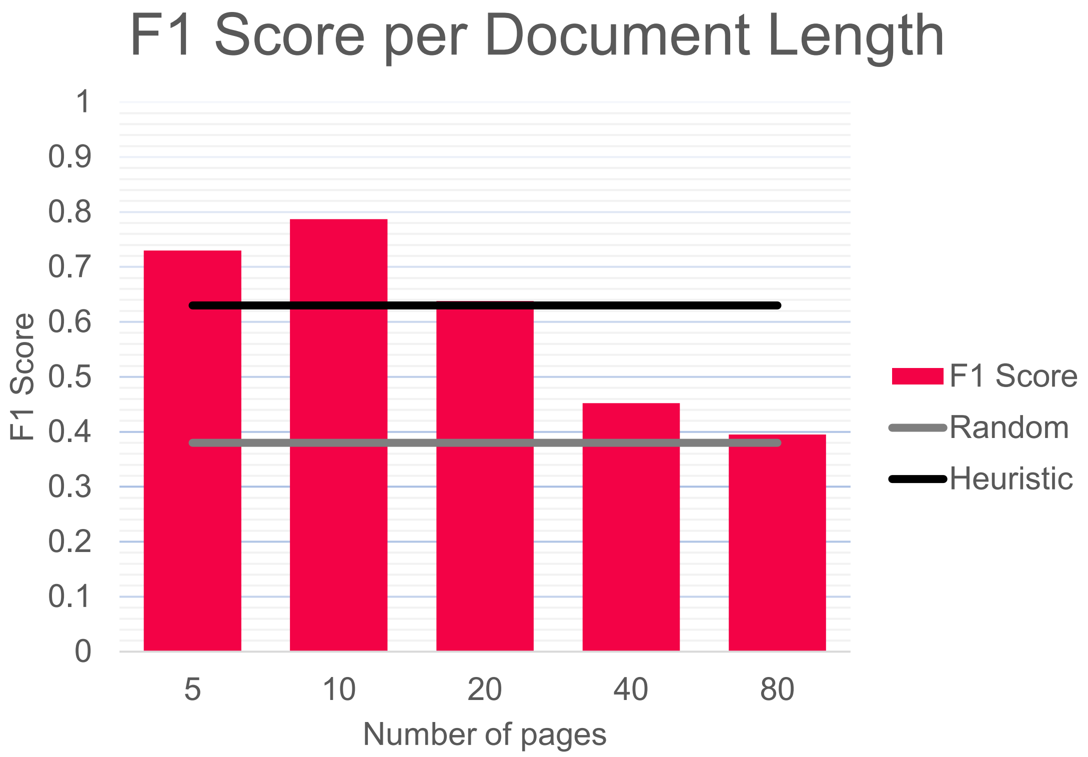
Post-processing

Merge/split rules improve boundary stability, especially for medium-length documents

Topic heterogeneity

Heterogeneous topics improve boundary precision; homogeneous content is harder to segment

Results — Chunking Behaviour



Sample size: 125 Test runs

Over-segmentation

F1 drops with document length, over-segmentation is the biggest problem at longer documents

LLM vs. Heuristic

LLM-based chunking outperforms heuristic baseline based on the TextTiling algorithm

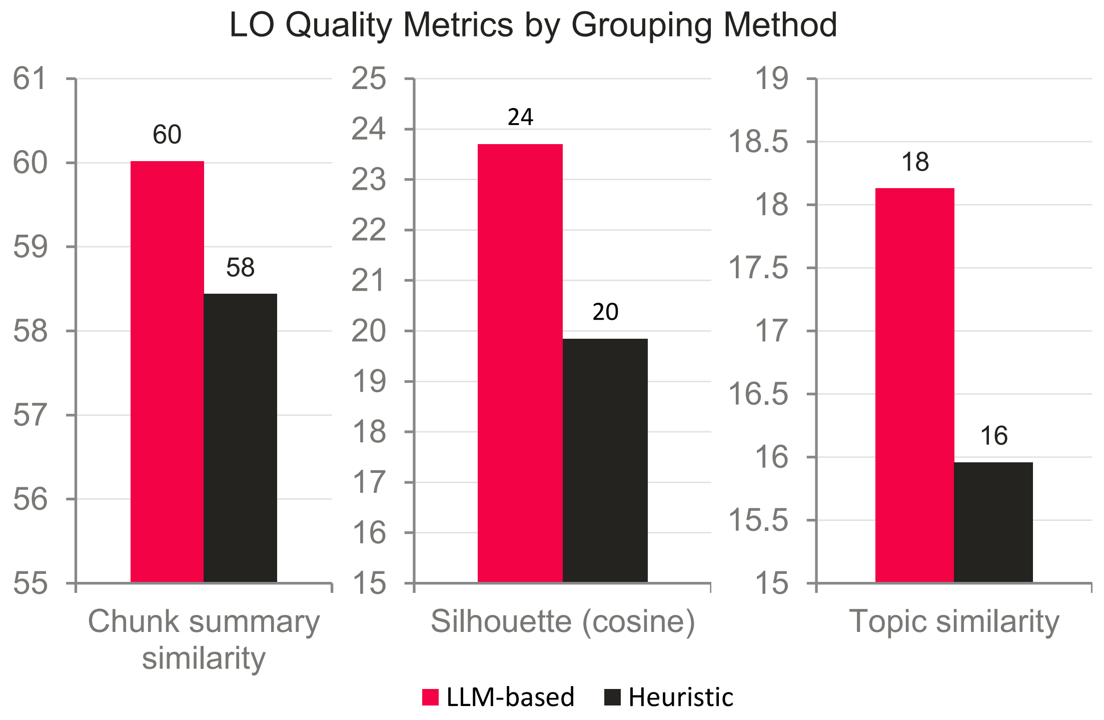
Post-processing

Merge/split rules improve boundary stability, especially for medium-length documents

Topic heterogeneity

Heterogeneous topics improve boundary precision; homogeneous content is harder to segment

Results — LO Grouping Quality



Sample size: 25 Test runs

Better Cluster Separation

Silhouette score improves **+20%** (24 vs 20), indicating clearer separation between different learning objective groups.

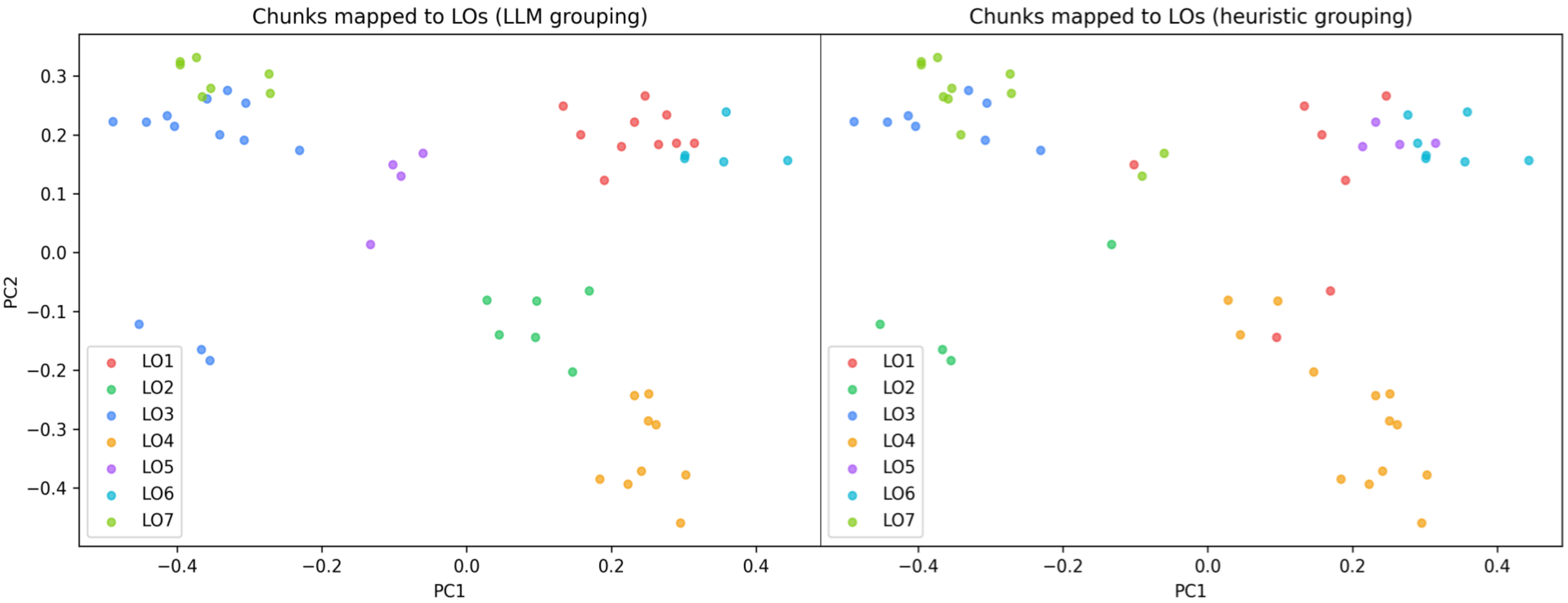
Higher Topic Coherence

Topic similarity increases **+14%** (18 vs 16), meaning chunks within a group share more consistent topics.

More Representative Summaries

Chunk summary similarity improves **+3%** (60 vs 58), showing summaries better reflect group content.

Results — Cluster Structure Visualization



Exploratory Expert Feedback

Qualitative evaluation with domain experts

Expert Observations

Chunking Quality

Chunks were generally perceived as topically coherent.

Experts occasionally observed very large chunks.

LO Structure

Generated LOs were considered useful for structuring course material.

Some chunk titles and LO labels were overly generic.

Usability Feedback

Navigation understandable after instructions.

Large PDFs caused slowdowns and missing loading feedback.

Ratings (1–5 scale)

Aspect	Mean
Chunk coherence	**4.0**
LO relevance	**4.5**
LO editing tools	3.5
Lesson quality	3.0
Quiz quality	2.5
UI clarity	**4.5**
Overall trust	3.5

Exploratory study with 2 completed expert responses. Results are descriptive and intended as qualitative feedback rather than statistical evaluation.

Limitations

Document length sensitivity

Segmentation quality decreases for very long documents.

Boundary density increases and global structure becomes harder to maintain as document length grows.

Synthetic benchmark scope

The benchmark is controlled and reproducible, but not a substitute for authentic noisy PDFs.

Real course material contains formatting noise, OCR errors, and inconsistent structure.

Structural, not pedagogical evaluation

Learning objectives were evaluated structurally, not pedagogically.

LOs were assessed using coherence and alignment metrics rather than expert educational evaluation.

Error propagation

Segmentation errors propagate into LO grouping and downstream lesson generation.

Key Conclusions

Research Question 1

LLM-based chunking is feasible and structurally meaningful under controlled evaluation.

Main limitation

Over-segmentation increases with document length and remains the key challenge.

Research Question 2

LLM-based LO grouping produces more coherent structures than heuristic baselines.

Overall system view

The prototype is promising for AI-assisted content structuring but requires further real-world validation.

Future Work

Real-world evaluation

Authentic educational PDFs with instructor-annotated boundaries to validate transfer from synthetic benchmarks.

Improved chunking

Fine-tuning, richer prompts, and layout-aware segmentation using document structure cues.

Human-centred validation

Instructor studies to evaluate pedagogical LO quality and downstream learning material usefulness.

Key takeaway: *Future work focuses on validating the pipeline in real educational settings and improving robustness of LLM-based segmentation.*

Thank You

Questions & Discussion

Thomas Plangger

Graz University of Technology · Institute of Human-Centred Computing

March 2026

gitlab.tugraz.at/ed-tech/2025/learning-objective-segmentation