RUHR-UNIVERSITÄT BOCHUM

AI-Driven Multistream Pipeline for Physics Lectures

Aleksei Mikhasenko[1], Mikhail Mikhasenko[2], Ilya Segal[2], Soraya Thiess[2]
[1]Bonn University, [2]Ruhr-University Bochum

Contact: s63amikh@uni-bonn.de

LECTURE NOTES
Goal: high-quality learning materials for students
Problem: combine several sources of information (recorded voice, student notes, and drawings)
Solution: use LLMs in several processing steps

Part 1: Initial Transcription and Raw Data Preparation

$0.5
AssemblyAI

Part 2: Processing of data

1. Step: create chunks using pre-trained model

Chonky
mirth/chonky_modernbert_large_1

Deepseek V3.1
$0.03

2. Step: remove unrelated content

3. Step: create blocks using weighted similarity

Snowflake Arctic
(**snowflake-arctic-embed2**)
in Ollama

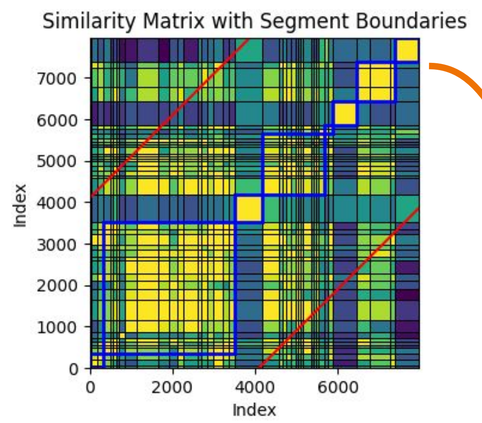
Blocks are created by maximizing this sum:

$$\left[\max_{\tau} \sum_{(i,j)\in\text{same block}} (M_{ij}-\tau), v_i v_j \quad \text{s.t.} \sum_{i\in\text{block}} v_i \le L_{\max.}\right]$$

- v_i: length (or weight) of chunk i
- M_{ij}: cosine similarity between chunks i, j, **divided by a configurable threshold** and **clipped at 1**

$$M_{ij} := \min\left(\frac{\cos(i,j)}{\text{threshold}}, 1\right)$$

- τ: similarity offset (configurable constant)
- $L_{\max}$: maximum total length allowed per block

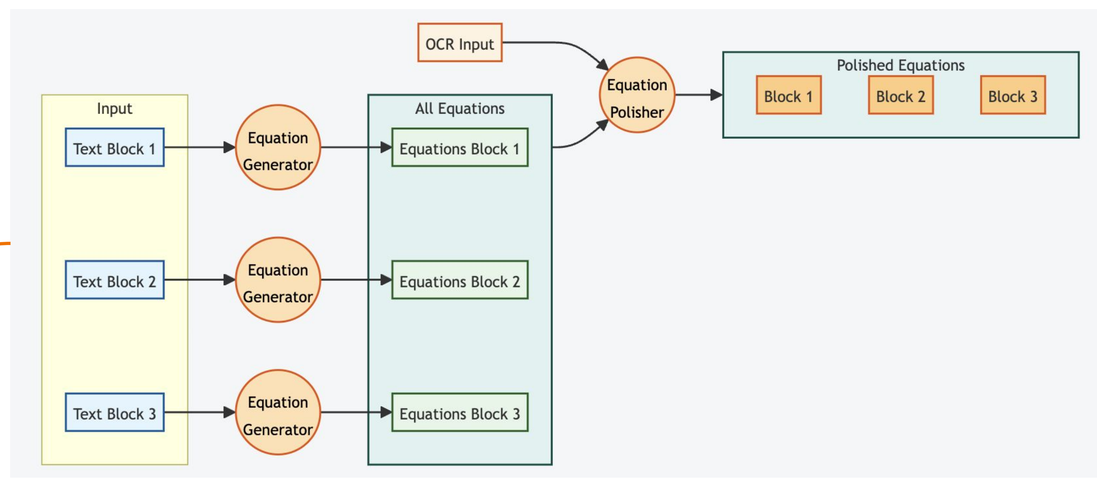


Visual segment separation

Configurable parameters

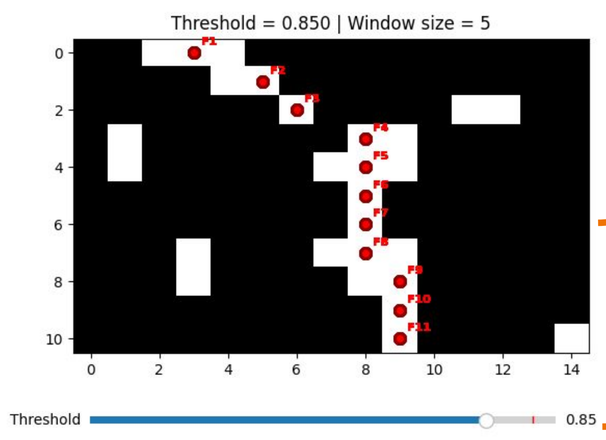
gpt-4.1-mini

4. Step: Create formulas from blocks



5. Step: Refine blocks with formulas

6. Step: Create formulas from blocks



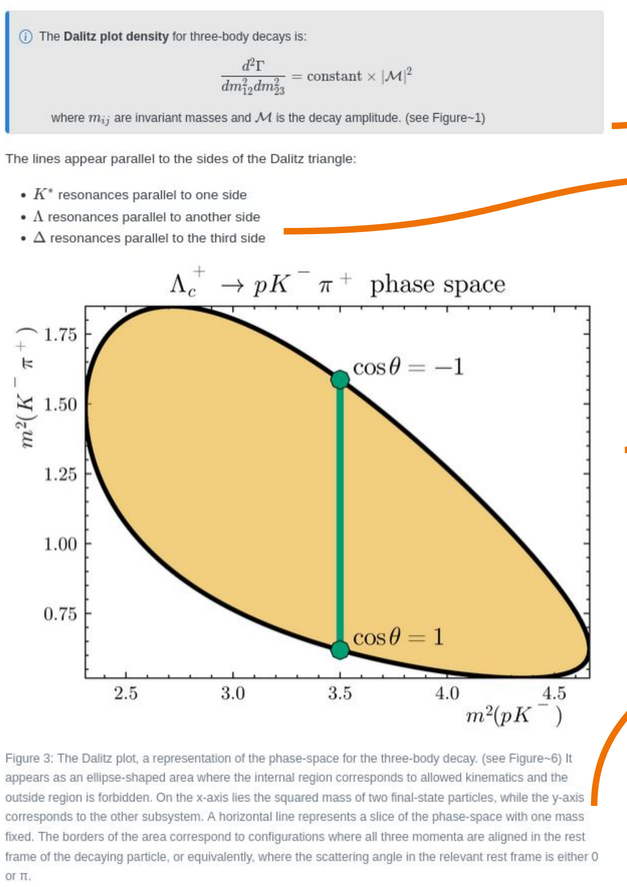
Images inserted starting from chunk with red dot in all white chunks

Threshold can be regulated

7. Step: Create titles for blocks

Deepseek V3.1

8. Step: Add markdown elements



Markdown notes

Markdown lists

gpt-4.1
$0.2

9. Step: Add images based on embedding similarity

Qwen-3 Embedding (8b)

Images in correct places

Labels for each image

10. Step: Save results to markdown

Part 3: Render to html and publish results

Check it out yourself !