# Data Literacy 2025 Project Report

**Ansel Cheung** [*1] **Alessio Villa** [*2] **Bartol Markovinović** [*3] **Martin Lopéz de Ipiña Munoz** [*4] **Niklas Abraham** [*5]

## Abstract

This project addresses the fundamental question of how cultural meaning evolves over time by quantitatively modeling seventy-five years of cinematic history through 200,000 film synopses embedded in a single static semantic space using the BGE-M3 model. Temporal change is measured by tracking the movement of genre centroids within this space—analyzing their velocity, acceleration, and curvature to distinguish continuous evolution from structural paradigm shifts. The framework provides a reproducible and data-driven foundation for cultural analytics, testing whether established linguistic laws of semantic drift extend to the domain of cinema.

## 1. Introduction

Motivate the problem, situation or topic you decided to work on. Describe why it matters (is it of societal, economic, scientific value?). Outline the rest of the paper (use references, e.g. to **??**: What kind of data you are working with, how you analyse it, and what kind of conclusion you reached. The point of the introduction is to make the reader want to read the rest of the paper.

## 2. Data and Methods

In this section, describe *what you did*. Roughly speaking, explain what data you worked with, how or from where it was collected, it's structure and size. Explain your analysis, and any specific choices you made in it. Depending on the nature of your project, you may focus more or less on certain aspects. If you collected data yourself, explain the collection process in detail. If you downloaded data from the net, show an exploratory analysis that builds intuition for the data, and shows that you know the data well. If you are doing a custom analysis, explain how it works and why it is the right choice. If you are using a standard tool, it may still help to briefly outline it. Cite relevant works. You can use the `\citep` and `\citet` commands for this purpose (**?**).

## 3. Results

In this section outline your results. At this point, you are just stating the outcome of your analysis. You can highlight important aspects ("we observe a significantly higher value of $x$ over $y$"), but leave interpretation and opinion to the next section. This section absoultely *must* include at least two figures.

## 4. Discussion & Conclusion

Use this section to briefly summarize the entire text. Highlight limitations and problems, but also make clear statements where they are possible and supported by the analysis.

[*] Equal contribution [1] Matrikelnummer 12345678, MSc Machine Learning [2] Matrikelnummer 12345678, MSc Computer Science [3] Matrikelnummer 7324790, MSc Machine Learning [4] Matrikelnummer 12345678, MSc Medical Informatics [5] Matrikelnummer 7307188, MSc Machine Learning. Correspondence to: Initials1 <first1.last1@uni-tuebingen.de>, Initials2 <first2.last2@uni-tuebingen.de>, Initials3 <first3.last3@uni-tuebingen.de>, Initials4 <first4.last4@uni-tuebingen.de>, Initials5 <niklas-sebastian.abraham@student.uni-tuebingen.de>.

*Project report for the "Data Literacy" course at the University of Tübingen, Winter 2025/26 (Module ML4201).* Style template based on the ICML style files 2025.

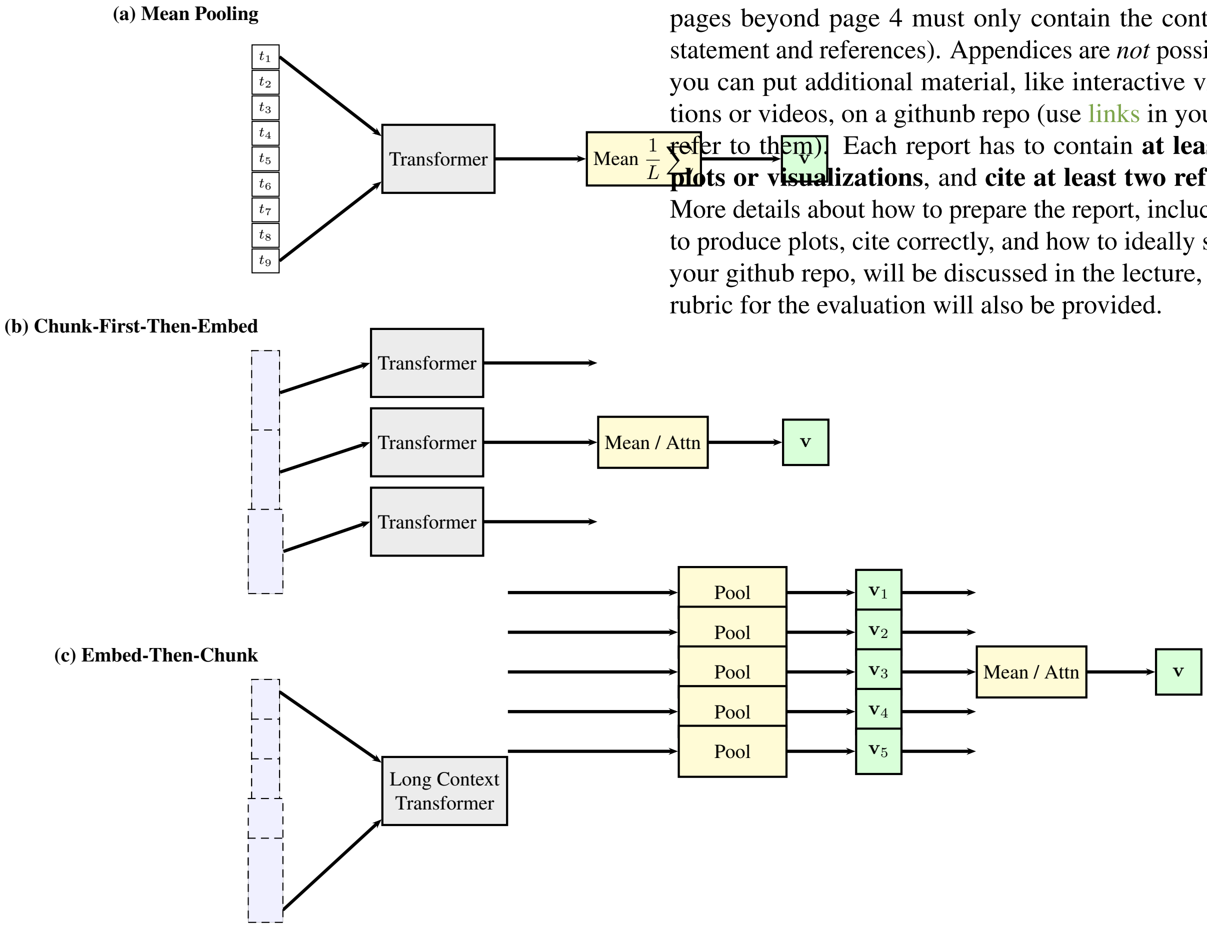

*Figure 1.* Three document-level pooling strategies: (a) Mean Pooling applies global averaging over all token embeddings, (b) Chunk-First-Then-Embed splits the document into chunks before encoding, then pools chunk embeddings, (c) Embed-Then-Chunk encodes the full document with a long-context transformer, then applies windowed pooling on the token embeddings before final aggregation.

## Contribution Statement

Explain here, in one sentence per person, what each group member contributed. For example, you could write: Max Mustermann collected and prepared data. Gabi Musterfrau and John Doe performed the data analysis. Jane Doe produced visualizations. All authors will jointly wrote the text of the report. Note that you, as a group, a collectively responsible for the report. Your contributions should be roughly equal in amount and difficulty.

## Notes

Your entire report has a **hard page limit of 4 pages** excluding references and the contribution statement. (I.e. any pages beyond page 4 must only contain the contribution statement and references). Appendices are *not* possible. But you can put additional material, like interactive visualizations or videos, on a githunb repo (use links in your pdf to refer to them). Each report has to contain **at least three plots or visualizations**, and **cite at least two references**. More details about how to prepare the report, inclucing how to produce plots, cite correctly, and how to ideally structure your github repo, will be discussed in the lecture, where a rubric for the evaluation will also be provided.