Optimizing AI-Driven Segmentation of Oral History: A Feedback-Loop Approach using Crowdsourced Validation in WO2Net

Feruza Bakhtiyorova & Dunya Boon
Digital Humanities in Practice 2026

Abstract.

Digital oral history collections require automated segmentation to be navigable, but current Artificial Intelligence (AI) models often struggle with the nuances of archival audio. The computational challenge lies in distinguishing historical narrative from technical "chatter" and administrative introductions. We analyzed 1250 crowdsourced validations to diagnose specific error patterns and implemented a refined prompt engineering strategy. Our evaluation compared baseline versus refined prompts on ten interview transcripts using structural metrics. The results show that our refined logic successfully removed non-narrative technical noise in 60% of cases and consolidated fragmented introductions, demonstrating that crowd feedback can be effectively used to optimize segmentation algorithms.

Keywords: Oral History; Segmentation; Large Language Models; Crowdsourcing; Prompt Engineering; Digital Humanities; GPT; World War 2; WO2Net.

1 Introduction

For this paper, we reason that the field of Digital Humanities contributes to the preservation of historical collections by digitizing and segmenting such oral history assemblies. Creating topic-specific fragments allows user audiences to connect with these narratives. The segmentation process for the WO2Net digital oral history collections is quickened by implementing AI-motivated automation, decreasing manual labor efforts. However, after manual checks, they found that AI is prone to produce non-specific outputs regarding topic selection. In order to improve specificity, we analyzed over 1250 crowdsourced user validations to define recurrent error patterns. Based on these error patterns, we adapted and refined the original prompt strategy with specific rules like trim chatter, specificity, and biographical noise. The trim chatter rule removes technical dialogue from interview segments like "band loopt" or "microfoon check". The specificity rule relates to titles and concepts where titles were marked to be too broad, like "Vertelt over de Oorlog" instead of "Arrest in Rotterdam", and concepts defaulted to broad themes like "Transport" instead of named entities like "Westerbork". The rule of biographical noise disregards interview fragments with solely biographical information, like names and dates of birth of interviewees. By adapting the original prompt strategy to include these rules, our goal is to improve AI-motivated segmentation output so that the process of automated segmentation is refined. This paper is a contribution to Digital Humanities research by quickening the automation process by way of implementing AI as opposed to manual labor in oral history data collections while preserving interview quality. This research also poses a publicly available coding strategy that can be reused and adapted by others to implement in other studies.

2 Related work

Automated video segmentation by use of AI is a strategy that is implemented in various fields of study, like skill assessment of surgeon performance (Permulla et al. 2023) and steganography (Lin et al., 2024), to name some examples. A study that handled an approach similar to our strategy using pre-trained models combined with human input by Gunnerli et al. (2024). They aim to develop a system for processing and analysing videos. Though the focus lies on advancing knowledge of human motion in dance through video analysis, their strategy to implement co-creative systems between humans and AI has proven to be successful in creating accurate motion capture and refined interpretation between the two. This approach is similar to the original strategy, where LLM-generated output was manually checked in order to refine the output, such as topic specifications. This combination of AI and human input is proven to be successful in both cases, but we argue that for our project, it would be possible to reduce human input and still be successful in creating refined outputs.

This is demonstrated to be possible by Ghosh et al. (2021). Instead of focusing on segmenting human motion, their research was centred around segmenting video lectures done by a video lecture augmentation system. These video lectures posed a challenge similar to one we encountered in handling eyewitness interviews. The video lectures failed to offer off-topic concepts, and our interviewees often deviated from topics, resulting in non-continuous topic presentation. Ghosh et al. (2021) handled this challenge with a system that can identify the off-topic topics and link them to other relevant video lectures to create a better understanding. In our project, we also handle the linking of topics and concepts to be grouped together, with the key difference being that linking and combining topics is done so in the same interview instead of linking to other sources. After implementing their system, a human-based evaluation was performed on the system, which indicated that the system was able to provide augmentation of video lectures by linking off-topic concepts. Though we note that their system for off-topic linking is based on manual models instead of generative AI, we suggest their results are an example of succeeding in creating a model for LLM systems to successfully link off-topic concepts within interviews.
Regarding related work of AI adaptations in a more Digital Humanities approach, Kabir et al. (2025) examine the benefits as well as the limitations of AI in qualitative research. They highlight that AI systems should not be considered a substitute for critical thinking and personal interpretations. We think it is important to include this take in our paper as well, since we do not aim to discard human input in our project, but for AI to be an established service to people.
Closer related to our work is that of Cocciolo (2025) on the use of ChatGPT to create oral history metadata in the form of summarized descriptions of interview transcripts. This particular approach focuses solely on summarizing data from transcripts through one specific prompt, therefore notably differentiating from our approach in creating a pipeline to extract specific information directly from video interviews.
The key difference that differentiates our work from that of others is that we provide a pragmatic middle ground where human input (crowdsourced validations) and AI automated segmentation collaborate, and where neither exceeds the other in the process. In addition, we acknowledge both elements' strengths and weaknesses during the process.

3 Problem statement/Dataset description

The oral history data collections consist of over 400 interviews, of which we worked with 355 enriched segment files. Video interviews are 'hours long', though we have no exact time specifications regarding the average duration. User validations are gained through a crowdsourcing platform, Streamlit, where users are presented with full-length interviews, transcripts, and an editable box with titles, concepts, start times, end times, fragment commentary, and a slide button to keep or remove the selected fragment. The original segments were created through AI, and users checked and improved these segments. The automatic segmenting of interviews is accomplished through a Python pipeline that contains the prompt engineering strategy. These segments are enriched with Second World War thesaurus concepts. We highly recommend readers inspect the original GitHub page in order to gain a better understanding of their methodology. For clarification, this is the original pipeline that we modified (Oorlogsbronnen, z.d.). By modifying the pipeline, our research question was formed.

RQ: Does refining prompt rules based on user feedback reduce rejection rates and improve segment quality in LLM-generated interview segmentation?

In our refined segmentation strategy, we faced computational challenges regarding filtering chatter while retaining quality and manually validating our output results. Generating segments with Large Language Models (LLMs), we had to manually check that interview fragments were not cut off in the middle of a sentence, the segments were conceptually rich to our three-rule standard, and no other conceptually rich information accidentally got excluded. In addition, feeding the prompts to the LLM is all done by hand, which takes time and results in a limiting sample size. The automation of the LLM's output is an aspect of our research that we are still working on. Regarding the humanities, we experienced difficulties with preserving the full narrative in interview segments. In discourse analysis, non-verbal cues represent an important part of effective communication (Mlakar et al., 2021). These non-verbal cues, like in-between silences, are at risk of being cut off or edited, therefore limiting analytical possibilities. This project is challenged in segmenting interviews and preserving the discourse as completely as possible.

4 Method/approach

We used a three-step approach: diagnosis, refinement, and evaluation. In order to gain more insights into our methodology, we recommend visiting the GitHub page we created and where we documented every step in greater detail (Ferufera, z.d.).

4.1 Preprocessing

We developed a utility script, consolidate_json_files.py, to prepare the raw data for analysis. This script aggregated the hundreds of individual JSON files stored in the source directories into two unified datasets: enriched_segments.json and segment_validations.json. While this step did not produce analysis results itself, it was helpful for optimizing the workflow. By consolidating the data once, we avoided the excessive processing load of individual files, ensuring that subsequent analysis scripts, specifically those used to detect mismatches, could run efficiently on a clean, unified dataset.

4.2 Solution approach and design

Based on the diagnosis, we implemented a refined prompt logic within refined_prompts.py. The refined prompt structure consists of seven strict segmentation rules. The first rule is ‘definition’, which defines segments as self-contained while focusing on one main theme. The second rule, ‘trim chatter’, excludes setup and technical language by changing the starting times to where the narrative in interviews begins. Rule three ‘merged introductions’ merges biographical data like names and dates of birth to be the first narrative segment to avoid short, incoherent fragments. The fourth rule, ‘topic shift’, ensures that a new segment only begins with a clear topic shift in order to keep fragments complete and not cut off. Rule five ‘segment length’ constrained segments to be between one and five minutes. The sixth rule, ‘overlap’, prohibits overlap between segments to ensure no captions are left unassigned with the exception of the exclusion of chatter by rule two. The seventh and last rule is ‘output’, which mandates the output format to be a valid JSON list of objects containing caption_indices.
While the baseline constraints ensure consistent formatting, segment length (1 to 5 minutes), and topic coherence (Rules1, 4-7), our optimization efforts focused specifically on two new interventions designed to address the high rejection rate:

- Anti-Chatter Protocol (Rule 2): We explicitly instructed the model to ignore technical setup phrases and anchor the start time to the first substantive narrative sentence.
- Introduction Merging (Rule 3): To prevent fragmentation, we forced the model to merge biographical data (Name, DOB) into the first thematic segment.

4.3 Evaluation setup

To test if this worked, we executed process_vtt_batch.py to run an A/B test on 10 full interview transcripts. We generated segments using the old logic (Original) and the new logic (Refined). The prompts were processed using the GPT-5.2 model (the latest free version available at the time of research) to generate the final JSON outputs.

We then used compare_results.py to automatically generate comparison reports for each interview. Figure 1 shows an example of such a report for interview 07_JKKV_Schelvis, highlighting the two key metrics: the “Start Index” shift (indicating chatter removal) and the Seg 1 Length” increase (indicating intro merging).

Figure 1: Example of automated comparison report (Interview: 07_JKKV_Schelvis)

```
1. EXECUTIVE SUMMARY
----------------------------------------
Original Segments: 23
Refined Segments:  24
Chatter Removal:   ✅ SUCCESS
Intro Merging:     ✅ SUCCESS

2. METRICS TABLE
--------------------------------------------------------------------------------
METRIC                         | ORIGINAL            | REFINED             | DELTA
--------------------------------------------------------------------------------
Start Index (Drift)            | 0                   | 2                   | +2
Seg 1 Length (Caps)            | 3                   | 14                  | x4.7
Avg Segment Length             | 21.8                | 14.8                | diff
--------------------------------------------------------------------------------
```

5 Results

The quantitative results from our A/B test showed clear improvements in structural coherence.

5.1 Removal of Technical Chatter

In 6 out of 10 interviews, the refined prompt produced a positive "Start Index Shift". This means the segment started later than before, successfully skipping the technical introduction. For example, in interview 22_stiso_Zeehandelaar, the start was shifted by 3 captions (about 10 seconds), cutting out the microphone check that users had previously complained about. In the other 4 cases, the shift was zero, which was correct because those interviews did not have any technical chatter to begin with.

5.2 Merging Introductions

The "Merge Intro" rule effectively stopped the creation of short fragments. As shown in Table 1, the length of the first segment increased noticeably in interviews where the original output was too short.

Table 1: Summary of Structural and Temporal Effects of Prompt Refinement

Interview ID	Chatter Removal	Intro Merged	Start Index Shift	Segment Count (Org → Ref)	First Segment Length (Org → Ref)	Overall Effect
07_JKKV_Schelvis	Yes	Yes	+2	23 → 24	3 → 14	Improved
22_stiso_Zeehandelaar	Yes	Yes	+3	17 → 6	6 → 37	Improved
GV_DeJager_Huijsman	Yes	Yes	+3	10 → 5	31 → 58	Improved
GV_DeJager_Drenth	Yes	No	+2	18 → 20	27 → 25	Partially improved
GV_DeJager_Schenk	Yes	No	+2	16 → 21	27 → 25	Partially improved
GV_DeJager_vanderBlom	Yes	Yes	+1	19 → 15	5 → 16	Improved
GV_NMKV_Sachsenhausen09	Yes	No	+2	13 → 19	20 → 18	Partially improved
GV_DeJager_BroederFrans	No	No	0	14 → 18	27 → 27	Stable
GV_DeJager_Ouwendijk	No	No	0	27 → 24	4 → 4	Stable
GV_DeJager_vanVliet	No	No	0	9 → 9	23 → 27	Stable

Table 1 summarizes the effects of the refined prompt rules across 10 interviews. Improvements are observed primarily in cases where technical setup or fragmented introductions were present, while interviews without such issues remain mostly unchanged. This indicated that the refined prompts act selectively, improving segmentation quality without overcorrecting or artificial forcing.

In the case of 07_JKKV_Schelvis (shown in Figure 1), the segments grew from 3 lines to 14 lines. This result is labeled "Merged" because the model successfully combined a broken introductory fragment with the main story. Results labeled "Expanded" (like GV_DeJager_Huijsman) indicate that an already valid segment grew larger where the original segmentation was already sufficient, confirming that the refined model intervenes selectively and avoids over-correcting valid structures. Cases labeled "Trimmed" (such as GV_DeJager_Drenth) show a slight reduction in length, confirming that the Atri-Chatter rule successfully removed the technical introduction while maintaining the stable segment boundary. Finally, cases labeled "Stable" indicate an interview where the original segmentation was already sufficient, confirming that the refined model intervenes selectively and avoids over-correcting valid structures.

6 Discussion and Conclusion

6.1 Discussion

This study is limited by the time constraints of four weeks. Consequently, while we fixed the structure (start times and merging), due to the time constraints, we did not solve the semantic issues. The validation data indicate that the automated solution did not fully meet user demands, specifically for high specificity in title edits and concept errors. Given more time, the same prompt-engineering logic could be applied to integrate a specific World War 2 thesaurus to fix these concept errors. Additionally, although our initial research design included a comparative analysis across multiple LLM architectures to test generalization, the short timeframe limited our evaluation to a single model (GPT-5.2). Another constraint is the comparative analysis of the ten interviews. This analysis had to be performed manually on a limited selection instead of being automated to check all interviews. As mentioned before the project is also in traction with humanities and capturing the full narrative, since in the segmentation progress non-verbal cues are at risk of being excluded from segments.
Finally, we learned significant lessons regarding the dynamics of interdisciplinary research. The collaboration between Artificial Intelligence and Communication Science required a transparent workflow where every methodology choice, from script to metric selection, was discussed and mutually validated. This constant dialogue ensured that the technical optimization remained aligned with the qualitative needs of the humanities. Furthermore, the strict project constraints emphasized the necessity of time management and realistic scoping, teaching us to prioritize high-impact structural interventions over broad, open-ended experiments.

6.2 Conclusion and takeaway

To answer the question: Does refining prompt rules based on user feedback reduce rejection rates and improve segment quality in LLM-generated interview segmentation? Our results show that we can lower the rejection rate by turning qualitative user complaints into specific prompt rules. By explicitly forbidding “chatter” and “short intros”, we aligned the AI output with human expectations. Analyzing crowdsource rejections is a powerful way to debug AI systems. Despite the temporal limitations, the study proves that even short-term interventions based on user feedback can improve archival processing pipelines. We consider that our findings provide strong indications that further developments allow for additional improvements in automated interview segmentation of oral history interviews to also fix conceptual errors and title edits. We would like to finish this paper with the takeaway that human input and LLM systems can be successfully combined to reduce time-consuming human efforts while holding on to specificity and complete interview segments.

7 Use of AI

In accordance with transparency, we would like to mention our implementation of AI in our project and final paper. Claude, a large language model developed by Anthropic (2025), was used for brainstorming purposes in generating digital humanities topics. ChatGPT (OpenAI, 2025) was used for text enhancement and translation in the methodology and results section of this paper. During the process of refining the code, ChatGPT was also used for debugging encountered errors.

References

Anthropic. (2025). Claude (version 4.5) [Large language model]. https://claude.ai

Cocciolo, A. (2025). Oral History Metadata and AI: A Study from an LGBTQ+ Archival Context. *Preservation Digital Technology & Culture*, *54*(1), 27–33. https://doi.org/10.1515/pdtc-2024-0054

Ferufera. (z.d.). *wo2-segmentation-optimization/02_optimization at main · ferufera/wo2-segmentation-optimization*. GitHub. https://github.com/ferufera/wo2-segmentation-optimization/tree/main/02_optimization

Ghosh, K., Nangi, S. R., Kanchugantla, Y., Rayapati, P. G., Bhowmick, P. K., & Goyal, P. (2021). Augmenting Video Lectures: Identifying Off-topic Concepts and Linking to Relevant Video Lecture Segments. *International Journal Of Artificial Intelligence in Education*, *32*(2), 382–412. https://doi.org/10.1007/s40593-021-00257-z

Gunerli, J. H. C., Deshpande, M., & Magerko, B. (2024). Video Segmentation Pipeline For Co-Creative AI Dance Application. *ACM Digital Library*, 1–5. https://doi.org/10.1145/3658852.3659085

Kabir, S. M. A., Ali, F., Ahmed, R. L., & Sulaiman-Hill, R. (2025). Exploring the Use of AI in Qualitative Data Analysis: Comparing Manual Processing with Avidnote for Theme Generation. *International Journal Of Qualitative Methods*, *24*. https://doi.org/10.1177/16094069251336810

Lin, Y., Luo, P., Zhang, Z., Liu, J., & Yang, X. (2024). AI‐generated video steganography based on semantic segmentation. *IET Image Processing*, *18*(11), 3042–3054. https://doi.org/10.1049/ipr2.13154

Mlakar, I., Rojc, M., Majhenič, S., & Verdonik, D. (2021). Discourse markers in relation tonon-verbal behavior. *Gesture*, *20*(1), 103–134. https://doi.org/10.1075/gest.20018.mla

Oorlogsbronnen. (z.d.). *GitHub - Oorlogsbronnen/wo2-oral-history-matching-pipeline: This project contains a Python pipeline for automatically segmenting World War II–related oral history interviews (in .vtt format) and enriching the segments with concepts from a WWII thesaurus.* GitHub. https://github.com/Oorlogsbronnen/wo2-oral-history-matching-pipeline/tree/main

OpenAI. (2025). ChatGPT (version GPT-5.2) [Large language model]. https://chat.openai.com

Perumalla, C., Kearse, L., Peven, M., Laufer, S., Goll, C., Wise, B., ... & Pugh, C. (2023). AI-based video segmentation: procedural steps or basic maneuvers?. *Journal of Surgical Research*, *283*, 500-506.