

# Paragraph Segmentation Revisited: Towards a Standard Task for Structuring Speech

Fabian Retkowski<sup>1</sup>, Alexander Waibel<sup>1,2</sup>

<sup>1</sup>Karlsruhe Institute of Technology, <sup>2</sup>Carnegie Mellon University

retkowski@kit.edu, waibel@cmu.edu

## Abstract

Automatic speech transcripts are often delivered as unstructured word streams that impede readability and repurposing. We recast paragraph segmentation as the missing structuring step and fill three gaps at the intersection of speech processing and text segmentation. First, we establish TED<sub>PARA</sub> (human-annotated TED talks) and YTSEG<sub>PARA</sub> (YouTube videos with synthetic labels) as the first benchmarks for the paragraph segmentation task. The benchmarks focus on the underexplored speech domain, where paragraph segmentation has traditionally not been part of post-processing, while also contributing to the wider text segmentation field, which still lacks robust and naturalistic benchmarks. Second, we propose a constrained-decoding formulation that lets large language models insert paragraph breaks while preserving the original transcript, enabling faithful, sentence-aligned evaluation. Third, we show that a compact model (MiniSeg) attains state-of-the-art accuracy and, when extended hierarchically, jointly predicts chapters and paragraphs with minimal computational cost. Together, our resources and methods establish paragraph segmentation as a standardized, practical task in speech processing.

**Keywords:** paragraph segmentation, text segmentation, transcript formatting

## 1. Introduction

Readability is a major concern for automatic speech recognition (ASR) transcripts, which are traditionally output as unstructured sequences of words, frequently lacking punctuation, casing, and higher-level organization such as sentence or paragraph boundaries (Jones et al., 2003; Shugrina, 2010; Tündik et al., 2018). While much research has focused on restoring sentence-level structure, paragraph segmentation is a less explored area. Paragraphs help users navigate and understand long-form speech such as lectures or meetings, where ideas unfold over extended spans (Lai et al., 2016). They also improve the usability and visual clarity of transcripts, avoiding the appearance of a dense, unreadable wall of text (Figure 1). Human evaluations show a preference for paragraph-segmented transcripts with breaks enhancing comprehension and perceived coherence (Pappu and Stent, 2015).

Despite its importance, paragraph segmentation remains underexplored in speech processing, in part due to the absence of standardized benchmarks and the scarcity of labeled data for spoken content. Unlike sentence boundary detection or punctuation restoration, paragraph segmentation lacks large-scale, curated datasets, making systematic evaluation and model development difficult.

To address this gap, we introduce TED<sub>PARA</sub> and YTSEG<sub>PARA</sub>, the first benchmarks for paragraph segmentation in speech, covering both TED talks and YouTube videos. Beyond that, these datasets fill a key void in the broader text segmentation area, where high-quality spoken-domain benchmarks remain scarce. We provide strong base-

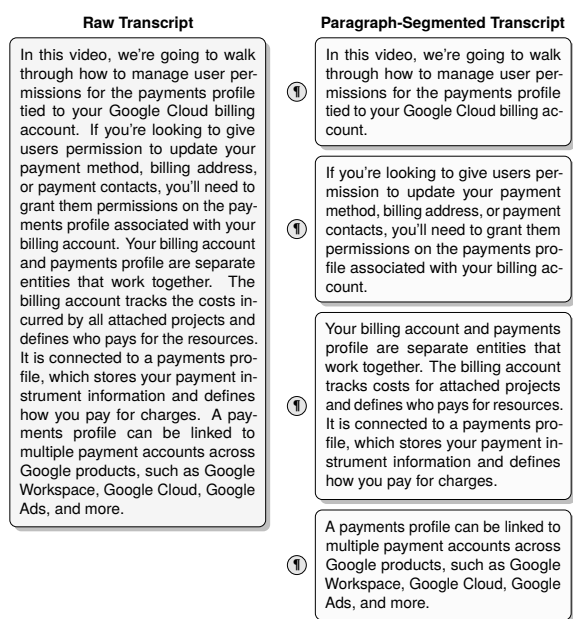


Figure 1: Paragraph segmentation turns an undifferentiated transcript into visually coherent paragraphs, aiding readability and navigation.

lines, including a compact model (MiniSeg) and LLM-based methods, and propose an efficient constrained decoding approach that inserts paragraph breaks while preserving transcript fidelity, essential for faithful evaluation. Using YTSEG<sub>PARA</sub>, we further extend MiniSeg to hierarchical modeling, enabling the joint prediction of chapter and paragraph boundaries, demonstrating that both levels can be learned efficiently within a shared framework. Evaluation combines automatic metrics with human

judgments, providing a comprehensive perspective on segmentation quality. Together, these contributions establish a foundation for treating paragraph segmentation as a standardized, measurable task in speech processing and highlight the practical value of our datasets for future research.

## 2. Related Work

**Paragraph Segmentation in Written Text.** Paragraph segmentation has largely been explored in the context of written text, where paragraph boundaries are typically already available. As a result, it is often treated as a secondary task, commonly used in self-supervised pretraining for sentence segmentation (Wicks and Post, 2021; Minixhofer et al., 2023; Frohmann et al., 2024), rather than as a primary research objective. Only a few recent studies have directly addressed paragraph segmentation as their main focus, typically within specific domains such as news articles and literary texts (Iikura et al., 2021; Zhuo et al., 2023; Yoo and Kim, 2024), but these efforts remain isolated without a standardized task definition or benchmark.

**Segmentation of Speech Transcripts.** Research on segmentation in speech data has been limited. Early work on video transcripts explored automatic paragraph segmentation strategies (Lai et al., 2016; Salimbajevs and Ikauniece, 2017; Lai et al., 2020), but these efforts did not produce reusable benchmarks, thereby limiting reproducibility and broader applicability. Recent work has focused on higher-level segmentation, targeting chapter segmentation in long-form spoken content with joint title generation (Ghazimatin et al., 2024; Retkowski and Waibel, 2024). While related, these approaches address a coarser segmentation granularity and involve broader objectives.

**Text Segmentation Benchmarks.** Current research in text segmentation is constrained by the scarcity of high-quality datasets. As noted by Glavaš et al. (2021), the field suffers from an “absence of large annotated datasets,” a limitation reflected in earlier work that often relied on small datasets or benchmarks constructed by concatenating unrelated snippets (e.g., Lukasik et al. 2020). A recent survey (Ghinassi et al., 2024) identifies this lack of suitable resources as the central challenge for progress. To date, only a few large-scale benchmarks exist, most prominently WIKI-727K (Koshorek et al., 2018) and more recently YTSeg (Retkowski and Waibel, 2024), which focus on topic- or chapter-level segmentation. In contrast, we introduce paragraph segmentation in spoken transcripts as a distinct task at a finer granularity. By establishing dedicated benchmarks for this setting,

we broaden the empirical landscape of segmentation research and enable more diverse and robust model evaluation across tasks and datasets under the shared framework of text segmentation.

**Constrained Decoding with LLMs.** Constrained decoding and structured output generation enable LLMs to produce outputs that follow specific formats or rules. Recent work has explored grammar-based and input-dependent constraints (Geng et al., 2023a,b). However, these works do not consider paragraph boundaries as a structured prediction problem.

## 3. Task Definition

*Paragraph segmentation* is a special case of *text segmentation* whose goal is to divide a text into paragraph units. While there is no single agreed-upon definition of what constitutes a *paragraph*, it is often described as a semantically or functionally coherent segment (Bolshakov and Gelbukh, 2001). At the same time, boundaries may also be introduced for stylistic reasons, considering discourse structure and rhetorical roles (Sporleder and Lapata, 2004), transitional and connective phrases (Zadrozny and Jensen, 1991; Lai et al., 2016), or length and readability (Yoo and Kim, 2024). Paragraph segmentation operates on a finer granularity compared to *topic segmentation*, which predicts higher-level topic shifts that also typically imply paragraph breaks (Filippova and Strube, 2006).

## 4. Dataset Construction

### 4.1. TEDPara

TEDPARA is derived from publicly available TED Talk transcripts, which include high-quality, human-annotated paragraph structure aligned with spoken presentations. We restrict our dataset to English transcripts only and collect all TED Talks listed on the official TED website as of May 13, 2024, spanning content published since February 2006. This results in an initial pool of 6,379 talks.

**Preprocessing.** We apply the following filtering steps. First, we remove all talks that lack a transcript, affecting 724 talks (12.8%). Next, we exclude talks that contain only a single paragraph, as they do not provide any paragraph boundary information; this step removes an additional 462 talks (7.2%). The final TEDPARA dataset contains 5,193 talks with multi-paragraph transcripts, which we randomly partitioned into training, validation, and testing splits; see Table 1 for details.

Split	# Talks	# Sent.	# Para.	Sent./Talk	Para./Talk	Sent./Para.	Breaks (%)
Train	4,154 (80%)	467,255	106,719	112.5	25.7	4.4 ± 4.0	22.0
Val	519 (10%)	60,257	13,697	116.1	26.4	4.4 ± 4.7	21.9
Test	520 (10%)	60,212	13,534	115.8	26.0	4.5 ± 4.1	21.6
Total	5,193	587,724	133,950	113.2	25.8	4.4 ± 4.1	21.9

Table 1: Data splits and dataset statistics for TED<sub>PARA</sub>

Dataset	Doc Len.	# Seg./Doc	Seg. Len.
TED <sub>PARA</sub>	113.2	25.8	4.4
YTSEG	196.2	9.12	21.5
WIKI-727K	57.6	3.48	13.6

Table 2: Segmentation granularity comparison across large-scale text segmentation datasets.

**Dataset Statistics.** Table 2 shows that TED<sub>PARA</sub> targets a finer segmentation level than large-scale benchmarks such as YTSEG and WIKI-727K. While these datasets contain fewer segments per document (9.12 and 3.48) with longer segments (21.5 and 13.6 sentences/segment), TED<sub>PARA</sub> has 25.8 paragraphs per talk with 4.4 sentences/paragraph on average. Together with its intermediate document length (113.2 sentences/talk), TED<sub>PARA</sub> complements existing benchmarks by expanding coverage in both granularity and document length.

**Release.** Due to licensing restrictions on TED content, we do not redistribute the data directly. Instead, we provide both the partitioned talk IDs as well as scripts for downloading and preprocessing the talks into a standardized format<sup>1</sup>, ensuring both reproducibility and legal compliance.

## 4.2. YTSegPara

To extend our task to more diverse speech content, we augment the existing YTSEG dataset (Retkowski and Waibel, 2024), which provides chapter annotations for structurally and topically diverse YouTube videos. The original dataset is limited to English-language videos with English transcripts, and the transcripts are derived from closed captions, which lack paragraph structure. We augment the dataset with paragraph-level annotations to produce a new dataset: YTSEG<sub>PARA</sub>.

Dataset	Para./Doc	Sent./Para.
TED <sub>PARA</sub>	25.8	4.4
YTSEG <sub>PARA</sub>	44.6	4.2

Table 3: Paragraph granularity comparison between TED<sub>PARA</sub> and YTSEG<sub>PARA</sub>.

<sup>1</sup><https://github.com/retkowski/tedseg>

**Augmentation.** Since manual paragraph annotation is infeasible at scale, we derive paragraph boundaries using the LLM-based constrained decoding method described in Section 5, using the LLaMA 3.1 70B model (Grattafiori et al., 2024).

**Dataset Statistics.** Table 3 compares the paragraph granularity of the two datasets. Both exhibit a similar paragraph density of  $\approx 4$  sentences per paragraph. However, YTSEG<sub>PARA</sub> contains nearly twice as many paragraphs per document (44.6 vs. 25.8), reflecting its longer documents.

**Utility and Scope.** Paragraph segmentation is a narrow problem that can be reduced to a sequence of binary decisions, making it well-suited for knowledge distillation into smaller, more efficient models (e.g., a Transformer encoder classifier) where inductive biases can effectively be imposed. Thus, the generated annotations serve a dual purpose: they provide training data for compact models and establish a benchmark for hierarchical segmentation, jointly predicting both chapters and paragraphs, in long-form spoken content.

**Release.** YTSEG<sub>PARA</sub> inherits the original dataset’s CC BY-NC-SA 4.0 license and is released with scripts and metadata<sup>2</sup>.

## 5. Paragraph Insertion with LLMs

We cast paragraph segmentation as a *constrained completion* task (Figure 2): at each sentence boundary the LLM may emit *only* (i) the punctuation token that ends the current sentence (“continue”), or (ii) punctuation tokens followed by the delimiter “\n\n” (“break”). Because the model cannot hallucinate arbitrary text, the original transcript is preserved verbatim and the paragraph structure emerges from a sequence of binary choices.

**Procedure.** Formally, let  $T$  be the transcript and  $S = (s_1, \dots, s_m)$  its sentence segmentation, obtained with NLTK’s sentence tokenizer (Bird et al.,

<sup>2</sup><https://github.com/retkowski/ytsegpara>

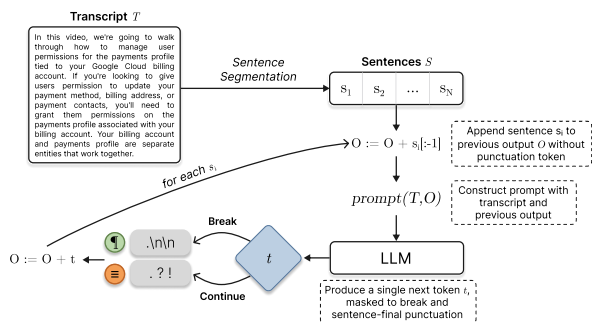


Figure 2: Conceptual illustration of the sentence-wise constrained decoding algorithm for paragraph insertion with LLMs. At each sentence boundary, the model performs a single forward pass to decide between two constrained actions: **continue** (emit standard punctuation) or **break** (emit punctuation followed by newlines).

2009). Before boundary  $i$  we hold the partially formatted output  $O$ , whose last token is the sentence-final punctuation  $p \in P$ , where  $P$  is the set of sentence-ending punctuation tokens (e.g., ., !, ?). We strip  $p$  (as a form of *token healing*; Lundberg and Ribeiro 2023) and build a prompt from the transcript  $T$  and the output prefix  $O$ . We tokenize this prompt as  $x = \tau(\text{prompt}(T, O))$ , where  $\tau$  is the model tokenizer, and query the LLM for the next-token distribution  $\mathbf{p} = \text{LLM}(x)$ , i.e.,  $\mathbf{p}[t] = p_{\text{LLM}}(t | x)$ . Let  $N$  be the set of allowed *break* tokens (punctuation tokens followed by the delimiter  $\backslash n \backslash n$ ). We mask all other tokens and compare:

$$p_{\text{punct}} = p_{\text{LLM}}(p | x), \quad (1)$$

$$p_{\text{break}} = \max_{t \in N} p_{\text{LLM}}(t | x). \quad (2)$$

If  $p_{\text{break}} > p_{\text{punct}}$  we append the most likely break token  $t^* = \arg \max_{t \in N} p_{\text{LLM}}(t | x)$  and set  $\pi_{i+1} = 1$  (paragraph break); otherwise we restore  $p$  and keep  $\pi_{i+1} = 0$ . Finally, we copy the next sentence  $s_{i+1}$  verbatim to  $O$ , adjusting whitespace to avoid consecutive spaces.

**Efficiency.** The loop performs exactly one forward pass of LLM per sentence boundary, for an  $O(m)$  runtime; far cheaper than the  $O(|\tau(T)|)$  generations a full re-write would need.

**Ensuring Comparable Evaluation.** Unconstrained decoding often modifies content or structure, producing outputs that differ from the reference transcript. Such deviations undermine evaluation reliability because automatic metrics in text segmentation including F1,  $P_k$  and Boundary Similarity depend on consistent sentence boundaries. If transcripts change, results become incomparable across systems.

**Applicability.** Importantly, constrained decoding is orthogonal to the choice of zero-shot versus fine-tuned inference. A task-adapted LLM benefits equally from the constrained formulation: it retains the efficiency of a single forward pass per boundary and guarantees verbatim transcript preservation.

**Section-Wise Processing.** When a dataset already provides coarser units such as chapters (as in YTSEG), we process each block independently. This ensures that predicted paragraph boundaries are coherent with the higher-level structure, while keeping inputs within the context window and enabling work on hierarchical segmentation.

**Generalization Across Tokenizers.** We observe two families of tokenization behavior with respect to paragraph delimiters. In the first, exemplified by LLaMA 3 (Grattafiori et al., 2024) and Qwen 2.5 (Yang et al., 2025) the tokenizer merges sentence-ending punctuation and the delimiter into a single compound token (e.g.,  $\backslash n \backslash n$ ). In the second, exemplified by Gemma 3 (Kamath et al., 2025), the delimiter  $\backslash n \backslash n$  is encoded as a separate token. The compound case is handled by the token-healing procedure described above. For the separate-token case, we can simplify it to a *next-word* variant: rather than stripping the final punctuation, we let the model score the first token of the following sentence and compare it against tokens that begin with  $\backslash n \backslash n$ . The decision rule is identical; the model chooses between continuing and breaking and only the token sets differ.

**Release.** For reproducibility and documentation, we publish the inference scripts with their decoding algorithm and prompts under a CC-BY 4.0 license<sup>3</sup>.

## 6. Experiments

### 6.1. Experimental Setup

Our experiments aim to establish strong baselines for our newly introduced benchmarks, TEDPARA and YTSEGPARGA, and to provide insights into the relationship between chapter segmentation and paragraph segmentation, the effectiveness of hierarchical segmentation, and the quality of LLM labeling used for pseudo-annotated data.

**Models.** In our experiments, we utilize the LLaMA 3 series, specifically LLaMA 3.1 8B, LLaMA 3.1 70B, and LLaMA 3.3 70B (Grattafiori et al., 2024) as examples for LLM-based paragraph segmentation, both in zero-shot and fine-tuned variants (LoRA).

<sup>3</sup><https://github.com/retkowski/paragraph-llm>

At the same time, we report results with MiniSeg (Retkowski and Waibel, 2024), as a model shown to have a strong performance in chapter segmentation of speech transcripts.

**Hierarchical Segmentation.** For experiments on YTSEG<sub>PARA</sub>, we extend MiniSeg to *hierarchical segmentation* by framing the segmentation on different levels (chapter-level and paragraph-level) as a *multi-class classification* problem.

**Metrics.** We report three metrics: F1 score, Boundary Similarity (BS; Fournier 2013), and  $P_k$  (Beeferman et al., 1999). F1 score is a classic classification metric that is increasingly used for text segmentation due to its interpretability. BS is a proposed and promising metric that captures graded boundary similarity.  $P_k$  is a well-established standard for evaluating segmentation tasks.

**Baselines.** Baselines include random and rule-based segmentations, multiple variants of MiniSeg trained with different pretraining and fine-tuning regimes, e.g., pretrained with WIKI-727K (Koshorek et al., 2018) as well as the LLaMA 3 series as an example for LLM-based segmentation.

**Random Baseline.** For TEDPara, we sample paragraph boundaries uniformly at random within each transcript, while assuming oracle knowledge of the reference number of paragraph breaks per document; this ensures the baseline matches the true break count but not their locations. For YTSegPara, we similarly sample chapter boundaries uniformly at random given the oracle number of chapters, then place paragraph breaks within each chapter span using the global paragraph-break rate.

**Rule-Based Baseline.** We include a simple deterministic baseline that inserts paragraph breaks at a fixed interval. Specifically, we place a boundary after every  $n$  sentences, where  $n$  is set to the global average sentences per paragraph in the corresponding split (rounded). For TEDPara, this gives  $n = 5$  for both validation (4.57) and test (4.63).

**Human Evaluation.** For human evaluation, we used two complementary methods. Pairwise comparisons were aggregated into ELO ratings, an increasingly common way to quantify relative preferences in a single measure (Boubdir et al., 2023; Chiang et al., 2024), while Likert-scale judgments (1–5) provided absolute quality assessments. This combination captures both relative and absolute perspectives on segmentation quality.

**Human Evaluation Methodology.** We conducted a two-part study on the TEDPara test split with 8 participants (4 per subtask), each completing 30 judgements. In Subtask 1, annotators performed randomized, blind A/B comparisons of paragraph segmentations with three response options (A, B, tie); presentation order and sampling were randomized to mitigate position bias and ensure balanced model coverage. In Subtask 2, annotators rated single segmentations on a 5-point Likert scale (1 = Poor, 5 = Excellent), using a similar balanced sampling procedure. Across both subtasks, we compared random and rule-based baselines, LLM-generated segmentations (with and without PBR), and human-annotated references.

**Token-Weighted LoRA Fine-Tuning.** Paragraph breaks are sparse at the sentence level and vanishingly so at the token level, where newline tokens are vastly outnumbered by content tokens. While adjusted losses have been explored for encoder-based paragraph segmentation (Iikura et al., 2021; Retkowski and Waibel, 2024), these operate on per-sentence classification. We investigate a simple analogue for autoregressive LLM fine-tuning: upweighting the newline token in the language modeling loss during training.

## 6.2. Results and Discussion

Model	Exact (↑)	+Whitespace (↑)	+Punct./Case (↑)	Len. ±5% (↑)
Naïve Decoding				
LLaMA 3.1 8B	0.59	0.80	0.83	0.99
LLaMA 3.1 70B	0.64	0.87	0.88	1.00
LLaMA 3.3 70B	0.63	0.89	0.91	1.00
Constrained Decoding				
LLaMA 3.1 8B	1.00	1.00	1.00	1.00
LLaMA 3.1 70B	1.00	1.00	1.00	1.00
LLaMA 3.3 70B	1.00	1.00	1.00	1.00

Table 4: Divergence of generated output from the original TEDPara transcript when using unconstrained decoding. Reported are proportions of outputs matching the input exactly (ignoring only paragraph breaks), then progressively relaxing constraints to ignore whitespace, punctuation/casing, and smaller length variations.

**Hallucination Risks with LLMs.** As demonstrated in Table 4, unconstrained decoding frequently introduces hallucinations, even under relaxed matching criteria. Although ignoring paragraph breaks, whitespace, and punctuation/casing improves match rates, divergence persists in a meaningful number of cases (ranging from 0.09 to 0.17, depending on the model). In the most lenient setting, which only requires outputs to be within 5% of the original length, LLaMA 3.1 8B still falls short of perfect fidelity (0.99), indicating occasional

severe hallucinations such as dropped text parts. These results strongly motivate the constrained decoding approach introduced in this paper, which not only improves efficiency but also ensures faithful preservation of the input transcript.

Model	Val			Test		
	F1 ( $\uparrow$ )	BS ( $\uparrow$ )	$P_k$ ( $\downarrow$ )	F1 ( $\uparrow$ )	BS ( $\uparrow$ )	$P_k$ ( $\downarrow$ )
<b>Zero-Shot</b>						
Random Baseline	16.8	15.9	49.7	17.1	15.8	49.6
Rule-Based Baseline	21.6	24.2	49.9	22.4	24.0	49.8
LLaMA 3.1 8B	36.6	30.5	38.7	33.9	28.3	40.3
LLaMA 3.1 70B	42.8	35.6	37.5	40.7	33.9	38.5
LLaMA 3.3 70B	37.5	30.1	37.7	37.1	29.3	38.2
MiniSeg (Wiki)	24.6	14.9	37.6	25.0	15.3	38.0
MiniSeg (Wiki) <sup>a</sup>	30.5	20.3	36.3	29.7	19.6	28.3
MiniSeg (Wiki $\rightarrow$ YT)	33.2	21.4	36.1	32.7	21.5	36.7
MiniSeg (Wiki $\rightarrow$ YT) <sup>a</sup>	45.7	32.6	32.4	43.4	30.9	33.2
<b>Zero-Shot + Paralinguistic Break Rule</b>						
LLaMA 3.1 8B + PBR <sup>b</sup>	51.7	44.3	32.9	49.9	42.5	34.2
LLaMA 3.1 70B + PBR <sup>b</sup>	55.5	47.6	32.4	54.0	46.7	33.1
LLaMA 3.3 70B + PBR <sup>b</sup>	52.2	43.7	32.3	52.5	43.5	32.5
<b>Fine-Tuned on TED<sub>PARA</sub></b>						
LLaMA 3.1 8B (LoRA; TED)	69.7	58.6	21.9	68.4	57.2	22.7
MiniSeg (TED)	67.3	56.1	24.3	67.3	56.3	24.0
MiniSeg (YT $\rightarrow$ TED)	69.8	60.2	22.4	70.6	61.2	21.6
MiniSeg (Wiki $\rightarrow$ TED)	70.6	60.7	21.8	71.2	61.5	21.0
MiniSeg (Wiki $\rightarrow$ YT $\rightarrow$ TED)	<b>72.1</b>	<b>62.2</b>	<b>21.0</b>	<b>72.7</b>	<b>63.2</b>	<b>20.1</b>

<sup>a</sup> Threshold tuned across partitions ( $\tau_{\text{val}}=0.264$ ,  $\tau_{\text{test}}=0.300$  for Wiki;  $\tau_{\text{val}}=0.257$ ,  $\tau_{\text{test}}=0.293$  for Wiki  $\rightarrow$  YT).

<sup>b</sup> Paralinguistic Break Rule (PBR): Additional, rule-based post-processing to insert paragraph breaks around standalone paralinguistic cues.

Table 5: Performance comparison of baselines for paragraph segmentation on TED<sub>PARA</sub>, using different approaches and (pre-)training strategies.

**From Chapters to Paragraphs.** The results in Table 5 highlight a strong connection between chapter and paragraph segmentation. The zero-shot performance of models trained on chapter-level data improves notably when the segmentation threshold is lowered, leading to more frequent segment predictions and outputs that better align with paragraph structure. Additionally, pretraining on data with higher-level segments such as WIKI-727K and YTSEG leads to strong results when fine-tuned on TED<sub>PARA</sub>, showing effective transfer of structural cues between domains and segmentation levels. Overall, the findings confirm that pretraining on related segmentation tasks significantly benefits paragraph segmentation.

**Paralinguistic Parentheticals.** We conducted a qualitative investigation to better understand the performance gap between the fine-tuned models and the LLM-based approaches. One consistent pattern we identified is that the TED<sub>PARA</sub> reference annotations reliably place paralinguistic parentheticals such as "(Laughter)" or "(Applause)" in their own paragraphs. In contrast, LLMs treat these as inline elements. This discrepancy introduces a systematic mismatch that is not due to a fundamental limitation of the LLMs, but rather the absence of a simple formatting rule. Once this rule is applied

by inserting paragraph boundaries before and after standalone paralinguistic cues, the performance gap narrows, as can be seen in Table 5.

Model	Paragraph Seg.			Chapter Seg.		
	F1 ( $\uparrow$ )	BS ( $\uparrow$ )	$P_k$ ( $\downarrow$ )	F1 ( $\uparrow$ )	BS ( $\uparrow$ )	$P_k$ ( $\downarrow$ )
Random Baseline	26.3	26.6	51.3	7.6	8.7	47.9
MiniSeg (Wiki $\rightarrow$ TED) <sup>a</sup>	35.7	32.2	43.6	—	—	—
MiniSeg (Wiki)	—	—	—	12.5	8.8	40.5
MiniSeg (Wiki $\rightarrow$ YT)	—	—	—	46.1	38.2	27.0
MiniSeg (YT)	47.9	42.2	33.4	43.6	32.5	28.8
MiniSeg (Wiki $\rightarrow$ YT)	50.5	44.5	32.2	43.8	35.7	28.2

<sup>a</sup> Predicted paragraph boundaries, scored within the oracle chapter spans.

■ Trained on either chapter or paragraph segmentation only.

■ Trained hierarchically on both paragraph and chapter segmentation.

Table 6: Performance comparison of baselines for hierarchical segmentation (consisting of paragraph segmentation and chapter segmentation) on YTSEG<sub>PARA</sub>, using different variants of MiniSeg.

**Efficient, Hierarchical Segmentation.** Table 6 shows that adding the paragraph task to MiniSeg results in only minimal impact on chapter-level performance: chapter F1 decreases slightly from 46.1 to 43.8, and  $P_k$  increases modestly from 27.0 to 28.2. This is notable given the increased inter-class confusion typically expected when expanding the label space. At the same time, the model produces a useful paragraph segmenter (paragraph F1 50.5). These results suggest that the existing parameter budget is sufficient to model both levels, yielding, to our knowledge, the first hierarchical segmentation model for speech and audiovisual transcripts.

Token Weight	Val				Test			
	P ( $\uparrow$ )	R ( $\uparrow$ )	F1 ( $\uparrow$ )	# Para.	P ( $\uparrow$ )	R ( $\uparrow$ )	F1 ( $\uparrow$ )	# Para.
1.0	74.9	65.1	69.7	21.7	74.1	63.5	68.4	20.6
1.5	70.8	68.8	69.8	24.4	69.6	67.9	68.8	23.7
2.0	67.2	72.5	69.7	27.0	66.2	73.5	69.7	27.3
Reference	—	—	—	26.4	—	—	—	26.0

Table 7: Effect of newline token weighting during LoRA fine-tuning on TED<sub>PARA</sub>. Higher weights increase the number of predicted segments, raising recall at the cost of precision.

**Token-Weighted Fine-Tuning.** In our LoRA fine-tuning setup, the model tends to slightly undersegment relative to the reference (Table 7). To explore whether this can be corrected, we experiment with upweighting the newline token in the language modeling loss during LoRA training. Increasing the weight encourages the model to predict paragraph breaks more frequently, effectively trading precision for recall. While this allows practitioners to calibrate the segmentation density to their needs, we find that adjusting the token weight does not improve the overall F1 score: gains in recall are offset by corresponding drops in precision.

Model	ELO Score	Model	Score
LLaMA 3.1 70B + PBR	1050.5	LLaMA 3.1 70B	3.64 ± 0.73
LLaMA 3.1 70B	1034.9	LLaMA 3.1 70B + PBR	3.55 ± 1.12
Reference	1015.9	Reference	3.50 ± 1.10
Rule-Based Baseline	1005.9	Rule-Based Baseline	3.30 ± 0.98
Random Baseline	892.8	Random Baseline	2.88 ± 1.32

(a) Pairwise preference evaluation using ELO. (b) Likert-scale ratings of segmentation quality.

Table 8: Human evaluation results on TEDPARA test data: (a) ELO scores from pairwise preference; (b) Likert-scale ratings (1 = Poor, 5 = Excellent).

**Human Validation of LLM Outputs.** As presented in Table 8, pairwise comparisons on TEDPARA yielded ELO ratings that place LLaMA 3.1 above the human reference and well above baselines, while Likert-scale judgments confirmed that its outputs are rated on par with references and consistently outperform rule-based or random baselines. These results are consistent with the automatic evaluation in Table 5, where LLM-based segmentation approaches the performance of supervised systems trained directly on reference segmentations. Together, these findings strengthen confidence that our approach produces paragraph structures comparable to human-authored segmentations and is suitable for use as a benchmark.

**Rule-Based Baseline Performance.** While rule-based segmentation performs poorly on automated metrics, its human evaluation scores are higher than expected given its simplicity. This likely reflects the visual structure it introduces: evenly spaced paragraph breaks reduce visual density and create a more readable layout. As Stark (1988) argued, paragraphing often serves stylistic functions rather than marking clear linguistic or semantic boundaries. The resulting visual separation can thus produce a sense of coherence and intentionality even when breaks are not meaningfully placed.

## 7. Conclusion

This work lays the foundation for treating paragraph segmentation as a standardized and measurable task in speech processing by introducing two complementary benchmarks: TEDPARA and YTSEGPARA. TEDPARA provides a high-quality, human-annotated reference grounded in formal spoken presentations, while YTSEGPARA covers a broad spectrum of real-world spoken content with synthetic labels generated via constrained LLM decoding. Together, these datasets capture a range of speech domains and conditions, forming a robust foundation for training and evaluation, not only for paragraph segmentation of speech transcripts but also in the broader research area of text segmentation, which has notoriously lacked benchmarks.

Our proposed constrained decoding method enables LLMs to efficiently and faithfully insert paragraphs without altering the original transcript. While fine-tuned models achieve stronger automatic scores, human evaluations rate the LLM outputs on par with or above human references, showing their potential as pseudo-labels for benchmarking. In addition, experiments demonstrate that paragraph and chapter segmentation can be modeled jointly with minimal performance trade-offs, enabling efficient, hierarchical structuring of speech transcripts. These contributions not only address a longstanding gap in transcript formatting but also support downstream tasks, including summarization and information retrieval, offering practical value for applications in education, accessibility, and knowledge management.

## 8. Limitations

While our benchmarks and methods advance paragraph segmentation for spoken content, several limitations remain. First, YTSEGPARA relies on synthetic labels generated via constrained decoding, which may not fully align with human judgment. However, our human evaluation provides encouraging evidence that LLM-generated segmentations broadly align with human preferences. Second, we observed a systematic stylistic mismatch between model and reference conventions, particularly in the handling of paralinguistic cues, which affects automatic metrics despite comparable perceived quality. Third, our datasets focus on structured, relatively clean speech from TED talks and YouTube videos, leaving out more challenging domains such as conversational meetings, where noise and disfluencies are more prevalent. Finally, our current approach operates solely on textual transcripts. While this simplifies processing and broadens applicability, it misses potentially useful prosodic cues from the audio signal, such as pauses, pitch, and intonation, that could further improve segmentation quality.

## 9. Acknowledgements

This research is supported by the project "How is AI Changing Science? Research in the Era of Learning Algorithms" (HiAICS), funded by the Volkswagen Foundation. In addition, we acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding pro-

gram. HoreKa is partly funded by the German Research Foundation (DFG).

## 10. Bibliographical References

- Doug Beeferman, Adam Berger, and John Lafferty. 1999. [Statistical Models for Text Segmentation](#). *Machine Learning*, 34(1):177–210.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*, 1st edition. O'Reilly, Beijing ; Cambridge [Mass.]. OCLC: ocn301885973.
- Igor A. Bolshakov and Alexander F. Gelbukh. 2001. [Text Segmentation into Paragraphs Based on Local Text Cohesion](#). In *Text, Speech and Dialogue, 4th International Conference, TSD 2001, Zelezna Ruda, Czech Republic, September 11-13, 2001, Proceedings*, volume 2166 of *Lecture Notes in Computer Science*, pages 158–166. Springer.
- Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. [Elo Uncovered: Robustness and Best Practices in Language Model Evaluation](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 339–352, Singapore. Association for Computational Linguistics.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning, ICLR'24*. JMLR.org. Place: Vienna, Austria.
- Katja Filippova and Michael Strube. 2006. [Using linguistically motivated features for paragraph boundary identification](#). In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 267–274, Sydney, Australia. Association for Computational Linguistics.
- Chris Fournier. 2013. [Evaluating Text Segmentation using Boundary Edit Distance](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1702–1712, Sofia, Bulgaria. Association for Computational Linguistics.
- Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. [Segment Any Text: A Universal Approach for Robust, Efficient and Adaptable Sentence Segmentation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941, Miami, Florida, USA. Association for Computational Linguistics.
- Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023a. [Grammar-Constrained Decoding for Structured NLP Tasks without Fine-tuning](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore. Association for Computational Linguistics.
- Saibo Geng, Martin Josifosky, Maxime Peyrard, and Robert West. 2023b. [Flexible Grammar-Based Constrained Decoding for Language Models](#). *CoRR*, abs/2305.13971. ArXiv: 2305.13971.
- Azin Ghazimatin, Ekaterina Garmash, Gustavo Penha, Kristen Sheets, Martin Achenbach, Oguz Semerci, Remi Galvez, Marcus Tannenber, Sahitya Mantravadi, Divya Narayanan, Ofeliya Kalaydzhyan, Douglas Cole, Ben Carterette, Ann Clifton, Paul N. Bennett, Claudia Hauff, and Mounia Lalmas. 2024. [PODTILE: Facilitating Podcast Episode Browsing with Auto-generated Chapters](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, pages 4487–4495, New York, NY, USA. Association for Computing Machinery. Event-place: Boise, ID, USA.
- Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. 2024. [Recent Trends in Linear Text Segmentation: A Survey](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3084–3095, Miami, Florida, USA. Association for Computational Linguistics.
- Goran Glavaš, Ananya Ganesh, and Swapna Somasundaran. 2021. [Training and Domain Adaptation for Supervised Text Segmentation](#). In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 110–116, Online. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, . . . , and Zhiyu Ma. 2024. [The Llama 3 Herd of Models](#). ArXiv:2407.21783 [cs].
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *ICLR*. OpenReview.net.

- Riku Iikura, Makoto Okada, and Naoki Mori. 2021. [Improving BERT with Focal Loss for Paragraph Segmentation of Novels](#). In *Distributed Computing and Artificial Intelligence, 17th International Conference*, pages 21–30, Cham. Springer International Publishing.
- Douglas A. Jones, Florian Wolf, Edward Gibson, Elliott Williams, Evelina Fedorenko, Douglas A. Reynolds, and Marc A. Zissman. 2003. [Measuring the readability of automatic speech-to-text transcripts](#). In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*, pages 1585–1588. ISCA.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Naveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, and Emilio Parisotto et al. 2025. [Gemma 3 technical report](#).
- Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. [Text Segmentation as a Supervised Learning Task](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.
- Catherine Lai, Mireia Farrús, and Johanna D. Moore. 2016. [Automatic Paragraph Segmentation with Lexical and Prosodic Features](#). In *Interspeech 2016*, pages 1034–1038. ISSN: 2958-1796.
- Catherine Lai, Mireia Farrús, and Johanna D. Moore. 2020. [Integrating lexical and prosodic features for automatic paragraph segmentation](#). *Speech Communication*, 121:44–57.
- Michał Łukasik, Boris Dachev, Kishore Papineni, and Gonçalo Simões. 2020. [Text Segmentation by Cross Segment Attention](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716, Online. Association for Computational Linguistics.
- Scott Lundberg and Marco Tulio Ribeiro. 2023. The art of prompt design: Prompt boundaries and token healing. *Towards Data Science (Medium)*.
- Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. [Where’s the Point? Self-Supervised Multilingual Punctuation-Agnostic Sentence Segmentation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. Association for Computational Linguistics.
- Aasish Pappu and Amanda Stent. 2015. [Automatic formatted transcripts for videos](#). In *Interspeech 2015*, pages 2514–2518. ISSN: 2958-1796.
- Fabian Retkowsky and Alexander Waibel. 2024. [From Text Segmentation to Smart Chaptering: A Novel Benchmark for Structuring Video Transcriptions](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian’s, Malta, March 17-22, 2024*, pages 406–419. Association for Computational Linguistics.
- Askars Salimbajevs and Indra Ikaunieca. 2017. [System for Speech Transcription and Post-Editing in Microsoft Word](#). In *Interspeech 2017*, pages 825–826. ISSN: 2958-1796.
- Maria Shugrina. 2010. [Formatting Time-Aligned ASR Transcripts for Readability](#). In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 198–206, Los Angeles, California. Association for Computational Linguistics.
- Caroline Sporleder and Mirella Lapata. 2004. [Automatic Paragraph Identification: A Study across Languages and Domains](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 72–79. ACL.
- Heather A. Stark. 1988. [What do paragraph markings do?](#) *Discourse Processes*, 11(3):275–303. Publisher: Routledge. eprint: <https://doi.org/10.1080/01638538809544704>.
- Máté Ákos Tündik, György Szaszák, Gábor Gostolya, and András Beke. 2018. [User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning](#). In *Interspeech 2018*, pages 2628–2632. ISSN: 2958-1796.
- Rachel Wicks and Matt Post. 2021. [A unified approach to sentence segmentation of punctuated text in many languages](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

*Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, and et al. 2025. [Qwen2.5 technical report](#).

Byunghwa Yoo and Kyung-Joong Kim. 2024. [Improving paragraph segmentation using BERT with additional information from probability density function modeling of segmentation distances](#). *Natural Language Processing Journal*, 6:100061.

Wlodek Zadrozny and Karen Jensen. 1991. Semantics of paragraphs. *Comput. Linguist.*, 17(2):171–209. Place: Cambridge, MA, USA Publisher: MIT Press.

Binggang Zhuo, Masaki Murata, and Qing Ma. 2023. [Auxiliary Loss for BERT-Based Paragraph Segmentation](#). *IEICE Transactions on Information and Systems*, E106.D(1):58–67.

## 11. Language Resource References

Koshorek, Omri and Cohen, Adir and Mor, Noam and Rotman, Michael and Berant, Jonathan. 2018. [Wiki-727K](#). TAD - The Center for AI & Data Science, Tel Aviv University. Distributed via GitHub.

Retkowski, Fabian and Waibel, Alexander. 2024. [YTSeg \(2024-07-25\)](#). Interactive Systems Lab, Karlsruhe Institute of Technology. Distributed via Hugging Face. PID <https://doi.org/10.57967/hf/1824>. License: CC-BY-NC-SA-4.0.

## A. Prompt Template

Figure 3 shows the prompt template used in both the naive and constrained decoding settings. We also employ prompt prefilling (i.e., including a stub assistant reply) to better guide the model toward generating paragraph-structured output.

## B. LLM Hyperparameters

**Greedy Decoding.** For all LLM-based inference, we apply *greedy decoding*. This applies to both the naive baseline as well as the constrained decoding setup. In the constrained decoding case, however, the output space at sentence boundaries is explicitly restricted to punctuation tokens only.

**Context Handling.** All LLaMA-based models used in our experiments support a context window of up to 128K tokens. This enables processing of long-form spoken content, including multi-hour YouTube transcripts from YTSEGPARA. While some transcripts exceed the model’s context limit, we employ chapter-wise processing, allowing all documents to be processed.

**LoRA Fine-Tuning.** We fine-tune LLaMA 3.1-8B-Instruct (Grattafiori et al., 2024) using LoRA (Hu et al., 2022) on the TEDPARA training set with the prompt template from Figure 3. To address the class imbalance between continuation and break tokens, we experiment with weighted cross-entropy, upweighting `\n\n` break tokens by a factor  $w \in \{1.0, 1.5, 2.0\}$  (see Table 7). The best checkpoint is selected by validation loss. All hyperparameters are listed in Table 9.

## C. MiniSeg Hyperparameters

For training MiniSeg on the TEDPARA and YTSEGPARA datasets, we largely follow the hyperparameter configuration established in the original work and implementation by Retkowski and Waibel (2024). The full set of hyperparameters used is summarized in Table 10. A key aspect of the training setup involves the weighting scheme for the weighted cross-entropy loss. For TEDPARA, we retain the original weighting of [1.2] as proposed. In contrast, YTSEGPARA involves a hierarchical segmentation task with three distinct classes, allowing for class-specific weighting. Based on validation experiments, a class weight configuration of [1, 1.5, 2] was found to be effective.

## D. Segmentation Evaluation

We utilize the `segeval`<sup>4</sup> library (Fournier, 2013) to calculate segmentation evaluation metrics, such as  $P_k$  and Boundary Similarity, using the default parameter configurations in both cases.

## E. Human Evaluation

We conducted a two-part human evaluation on the TEDPara test dataset with 8 participants (4 per subtask; 30 trials each), comparing random and rule-based baselines, LLM-generated segmentations (with and without PBR), and human-annotated references. In Subtask 1, annotators performed randomized, blind pairwise A/B comparisons of segmentations with three response options: A, B, or tie; to mitigate position bias, presentation order was randomized, and model-text pairs were sampled online using inverse-frequency weighting to avoid per-participant repeats and ensure balanced coverage. In Subtask 2, participants rated individual segmentations on a 5-point Likert scale, with each trial assigned by a balanced sampler that inversely weighted overrepresented models and prioritized unseen texts, again ensuring broad and nonredundant evaluation coverage.

---

<sup>4</sup><https://segeval.readthedocs.io/>

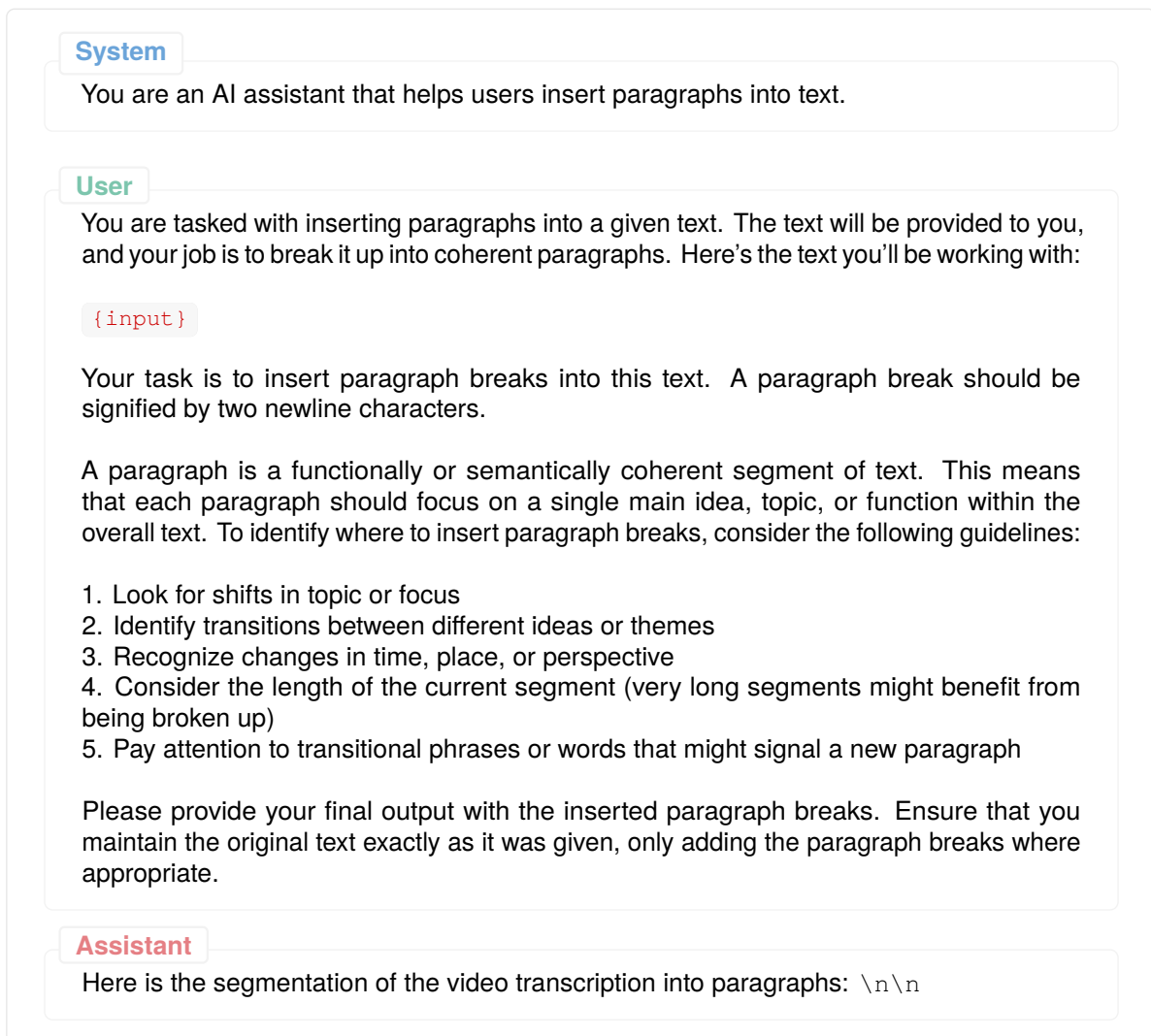


Figure 3: Prompt Template for Paragraph Insertion with LLMs

Hyperparameter	Value
Base Model	LLaMA 3.1-8B-Instruct
LoRA Rank ( $r$ )	16
LoRA Alpha ( $\alpha$ )	32
LoRA Dropout	0.05
LoRA Target	All linear layers
Max Sequence Length	8,192
Epochs	3
Effective Batch Size	32
Learning Rate	$1 \times 10^{-4}$
LR Schedule	Cosine
Warmup Ratio	0.1
Precision	bfloat16
Checkpoint Selection	Best validation loss
Break Token Weight ( $w$ )	{1.0, 1.5, 2.0}

Table 9: LoRA fine-tuning hyperparameters.

Hyperparameter	TED <sub>PARA</sub>	YTSEG <sub>PARA</sub>
Sentence Encoder		all-MiniLM-L12-v2
Loss Function	Weighted Binary Cross-Entropy	
Learning Rate		$2.5 \times 10^{-5}$
Batch Size		115,000 Tokens
Epochs		15
Learning Rate Schedule		Cosine
Optimizer		AdamW
Dropout Rate		0.1
Gradient Sampling Rate		0.5
Cross-Entropy Class Weights	[1.2]	[1, 1.5, 2]

Table 10: MiniSeg Hyperparameters for Training on TED<sub>PARA</sub> and YTSEG<sub>PARA</sub>