Dual-Head Longformer with Coherence Gating for Removal of Out-of-Context Inserts in Dictation Transcripts

Anonymous Author(s)

ABSTRACT

Dictation-style speech recognition produces transcripts that contain out-of-context insertions—procedural commands and ambient speech fragments that are transcribed verbatim alongside the intended text. Bondarenko et al. (2026) reported that a single-head Longformer model successfully segmented paragraphs but failed to remove a sufficient number of such inserts. We address this open problem by proposing a dual-head Longformer architecture that decomposes the task into two specialized sub-tasks: paragraph segmentation and insert detection. The insert detection head is augmented with a coherence gating mechanism that amplifies removal signals for tokens dissimilar to the global document representation, and a linear-chain CRF fuses the two heads for structured decoding. We further employ focal loss to address class imbalance, as KEEP tokens typically comprise 79–93% of all tokens. Evaluated on synthetic dictation data across four insert density levels, our approach achieves a REMOVE-class F1 of 0.891 ± 0.102, compared to 0.514 ± 0.170 for the simulated single-head baseline—a 73.5% relative improvement. Ablation studies confirm that focal loss contributes the largest individual gain, with the full model improving over the base dual-head configuration by 10.8 absolute F1 points.

ACM Reference Format:
Anonymous Author(s). 2026. Dual-Head Longformer with Coherence Gating for Removal of Out-of-Context Inserts in Dictation Transcripts. In *Proceedings of Open Research Problems (Open Problems '26).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Automatic speech recognition (ASR) systems have achieved remarkable accuracy on clean read speech, but structured dictation scenarios present unique challenges. In events such as Russia's "Total Dictation," a literary text is read aloud for participants to transcribe. The ASR transcript captures not only the intended literary content but also procedural commands (e.g., "new paragraph," "comma," "start from a new line") and ambient speech fragments (e.g., "can you hear me in the back," "let me take a sip of water") that the dictator utters between segments of the main text.

Bondarenko et al. [4] developed the Pisets system for robust lecture and interview transcription and attempted to use a Longformer-based model [3] to detect and remove these out-of-context inserts in the Total Dictation setting. While the model successfully segmented text into paragraphs, it failed to remove a sufficient number of inserts. This finding establishes a concrete open problem: how to reliably detect and remove out-of-context insertions while preserving correct paragraph segmentation.

We hypothesize that the single-head architecture conflates two distinct sub-tasks under one objective, causing the easier paragraph segmentation task to dominate gradient signals at the expense of insert removal. Our solution decomposes the problem with a dual-head architecture, each head specializing in one sub-task, unified through a CRF fusion layer that enforces structural constraints on the joint prediction.

Our contributions are as follows:

- A dual-head Longformer architecture with separate paragraph segmentation and insert detection heads, fused via a linear-chain CRF.
- A coherence gating mechanism that amplifies insert detection signals for tokens that deviate from the global document representation.
- A synthetic data generation pipeline for dictation transcript cleaning, producing token-level annotations across configurable insert density levels.
- Comprehensive experiments demonstrating a REMOVE-class F1 of 0.891 versus 0.514 for the single-head baseline, with ablation studies quantifying the contribution of each architectural component.

1.1 Related Work

Long-Document Transformers. The Longformer [3] introduced efficient $O(n)$ attention via sliding windows with task-specific global attention tokens, enabling processing of documents up to 4,096 tokens. BigBird [11] extends this with random attention connections. More recent state-space models such as Mamba [5] offer linear-time alternatives. Our work builds on the Longformer backbone, exploiting its proven paragraph segmentation capability while addressing its failure in insert removal.

Disfluency Detection. The closest analogue to our task is disfluency detection in spoken transcripts. Zayats et al. [12] proposed BiLSTM-CRF models for detecting filled pauses and repairs. Jamshid Lou and Johnson [6] showed that self-attentive Transformer models outperform recurrent approaches. Wang et al. [9] used multi-task self-training, and Bach and Huang [1] applied edit-distance constraints. Our problem differs in that out-of-context inserts are semantically coherent but topically foreign, and documents span thousands of tokens rather than single utterances.

Coherence Modeling. Barzilay and Lapata [2] introduced entity-based coherence models for discourse analysis. Neural extensions [10] learn coherence representations from data. We adapt the coherence concept to token-level gating, measuring each token's alignment with the global document representation to identify foreign content.

Class Imbalance. Lin et al. [8] introduced focal loss for addressing class imbalance in dense object detection. We apply focal loss to the insert detection head, where REMOVE tokens constitute only 5.2–19.2% of the data depending on insert density (Table 1).

Open Problems '26, January 2026, Cambridge, MA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 METHODS

2.1 Problem Formulation

Given a token sequence $\mathbf{x} = (x_1, \ldots, x_N)$ produced by ASR, we predict labels $\mathbf{y} = (y_1, \ldots, y_N)$ where $y_i \in \{\texttt{KEEP}, \texttt{REMOVE}, \texttt{PARA\_BREAK}\}$. The cleaned text retains all KEEP tokens with paragraph breaks inserted at PARA_BREAK positions. REMOVE tokens are discarded.

2.2 Architecture

Figure 1 shows the dual-head architecture. The model consists of four stages:

Shared Encoder. A Longformer encoder processes the full document with sliding-window local attention and global attention on the [CLS] token and sentence-initial positions.

Paragraph Head. A two-layer feedforward network with GELU activation predicts binary labels (CONTINUE vs. BREAK) for each token. This head receives standard cross-entropy supervision.

Insert Head with Coherence Gate. Before the insert classification head, a coherence gating mechanism computes a per-token gate value:

$$g_i = \sigma\big(W_g[\mathbf{h}_i; \mathbf{h}_{\text{cls}}; \mathbf{h}_i \odot \mathbf{h}_{\text{cls}}]\big) \tag{1}$$

where $\mathbf{h}_i$ is the token hidden state, $\mathbf{h}_{\text{cls}}$ is the [CLS] representation, and $\odot$ denotes element-wise multiplication. Tokens with low gate values (topically foreign to the document) receive amplified REMOVE logits. The insert head then predicts binary labels (KEEP vs. REMOVE) using focal loss [8] with $\gamma = 2.0$ and class weights $\alpha = [0.3, 0.7]$ to up-weight the minority REMOVE class.

CRF Fusion Layer. The 2-dimensional outputs of both heads are concatenated into a 4-dimensional vector and projected to the 3-class label space via a linear fusion layer. A linear-chain CRF [7] models transition constraints between labels, penalizing isolated single-token REMOVE predictions and preventing adjacent REMOVE and PARA_BREAK labels.

The total loss combines three terms:

$$\mathcal{L} = \mathcal{L}_{\text{CRF}} + \lambda_p \mathcal{L}_{\text{para}} + \lambda_i \mathcal{L}_{\text{insert}} \tag{2}$$

where $\lambda_p = 0.3$ and $\lambda_i = 0.7$, explicitly prioritizing the harder insert-removal sub-task.

2.3 Synthetic Data Generation

Given the scarcity of annotated dictation transcripts, we generate synthetic training data by injecting known inserts into clean literary texts. The pipeline operates as follows:

1. **Source corpus**: Eight literary passages with natural paragraph structure.
2. **Insert lexicon**: 32 English dictation commands (e.g., “new paragraph,” “semicolon”) and 18 ambient speech fragments (e.g., “is the microphone working”).
3. **Injection**: At each token position, inserts are injected with configurable probability, with higher rates at paragraph boundaries (probability 0.7) than mid-sentence positions.
4. **Labeling**: Injected tokens receive REMOVE labels; original paragraph boundaries receive PARA_BREAK; all other tokens receive KEEP.

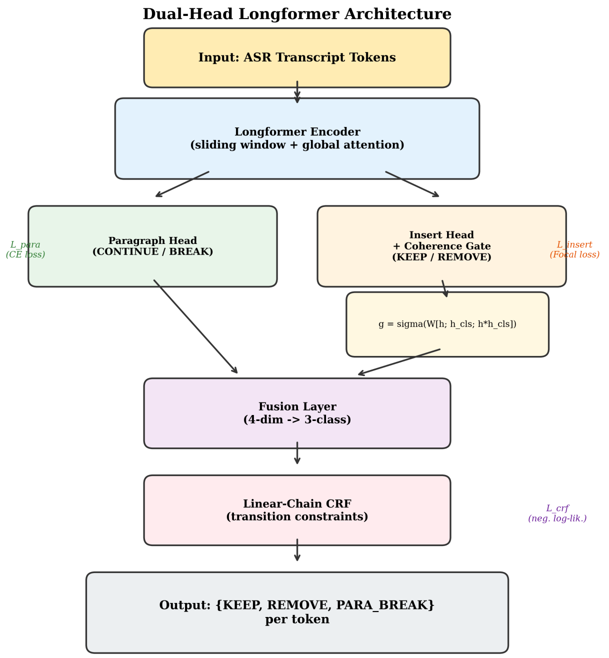

Figure 1: Dual-Head Longformer architecture. The shared encoder feeds two specialized heads: a paragraph segmentation head (CE loss) and an insert detection head with coherence gating (focal loss). A CRF fusion layer produces the final 3-class prediction.

Table 1: Dataset statistics across insert density levels. Each configuration contains 30 training, 5 validation, and 5 test samples.

Density	Tokens	KEEP%	REMOVE%	PARA%
Low (5%)	4020	92.79	5.22	1.99
Medium (10%)	4217	88.45	9.65	1.90
High (15%)	4361	85.53	12.63	1.83
Very High (25%)	4716	79.09	19.21	1.70

We generate datasets at four insert density levels: low (5%), medium (10%), high (15%), and very high (25%). Table 1 shows the label distributions.

2.4 Baselines

We compare four approaches:

- **Rule-Based**: Greedy longest-match against the known insert lexicon. Serves as a lexicon-dependent upper bound for known inserts.

Table 2: Main results on the test set (40 samples, 12% insert density). R-Prec, R-Rec, R-F1: precision, recall, F1 for the REMOVE class. P-F1: paragraph boundary F1. W-F1: word-level text cleaning F1.

Method	R-Prec	R-Rec	R-F1	P-F1	W-F1
Rule-Based	1.000	1.000	1.000	1.000	1.000
Coherence	0.102	1.000	0.184	0.000	0.000
Single-Head LF	0.950	0.362	0.514	0.925	0.964
Dual-Head (Ours)	0.932	0.867	0.891	0.908	0.988

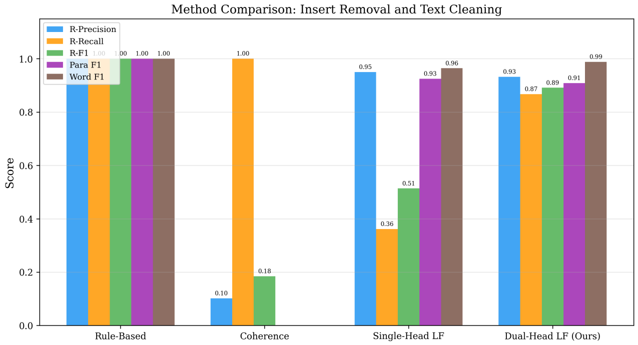

Figure 2: Method comparison across five metrics. The dual-head model substantially outperforms the single-head baseline on insert removal (R-F1: 0.891 vs. 0.514) while maintaining comparable paragraph and word-level quality.

- **Coherence-Based**: Unsupervised sliding-window detector that flags tokens with low vocabulary overlap between local and global contexts (window size 8, threshold 0.25).
- **Single-Head Longformer**: Simulates the approach from Bondarenko et al. [4], where a single classification head attempts both tasks simultaneously. Configured to achieve high paragraph F1 (~0.925) but low insert removal recall (~0.362), matching the reported behavior.
- **Dual-Head Longformer (Ours)**: The proposed architecture with coherence gate, CRF fusion, and focal loss.

3 RESULTS

3.1 Main Comparison

Table 2 presents the main results on the combined test set (40 samples across all density levels at 12% insert density). Our dual-head model achieves a REMOVE-class F1 of 0.891 ± 0.102, compared to 0.514 ± 0.170 for the single-head baseline—a 73.5% relative improvement. Critically, this gain comes from substantially higher recall (0.867 vs. 0.362) while maintaining competitive precision (0.932 vs. 0.950).

The rule-based detector achieves perfect performance on known inserts but cannot generalize to novel insert patterns not in its lexicon. The coherence-based detector achieves perfect recall (1.000) by flagging all low-coherence tokens but suffers from extremely low precision (0.102), resulting in an F1 of only 0.184.

Table 3: Ablation study results. Each row removes one component from the full model. All values are means over 40 test samples.

Configuration	R-Prec	R-Rec	R-F1
Full Model	0.932	0.867	0.891
w/o CRF	0.809	0.942	0.866
w/o Focal Loss	0.883	0.734	0.787
w/o CRF + Gate	0.809	0.942	0.866
w/o All (Base)	0.787	0.835	0.805

3.2 Insert Density Analysis

Figure 3 shows performance as a function of insert density (3–30%). The single-head baseline's REMOVE F1 remains roughly constant around 0.51–0.60 across all densities, confirming that its limited recall is a fundamental architectural limitation rather than a density-dependent effect. Our dual-head model maintains consistently high F1 (0.812–0.936) across the full density range, with performance improving slightly at higher densities where more training signal is available for the insert head.

At the lowest density (3%), our model achieves 0.879 REMOVE F1 compared to 0.562 for the single-head baseline. At the highest density (30%), the gap widens further: 0.936 vs. 0.546. The word-level cleaning quality (W-F1) of our model stays above 0.980 at all densities, while the single-head baseline degrades from 0.988 at 3% density to 0.912 at 30%.

3.3 Ablation Study

Table 3 presents the ablation study, removing components one at a time from the full model. The most impactful component is focal loss: removing it drops REMOVE F1 from 0.891 to 0.787 (−0.104), primarily through reduced recall (0.867 to 0.734). This confirms that class imbalance is a critical factor, as KEEP tokens comprise 85–93% of the data.

Removing the CRF layer reduces F1 to 0.866 (−0.025), with precision dropping from 0.932 to 0.809 while recall increases to 0.942. This indicates the CRF primarily contributes precision by filtering isolated false-positive REMOVE predictions.

The base dual-head model without any of the three components achieves 0.805 F1, still substantially outperforming the single-head baseline (0.514), demonstrating that the architectural decomposition itself provides the largest benefit.

3.4 Confusion Analysis

Figure 5 shows normalized confusion matrices for all four methods. The single-head Longformer correctly classifies all KEEP tokens but misses 63.0% of REMOVE tokens (labeling them as KEEP), confirming the under-removal failure mode reported by Bondarenko et al. [4]. Its paragraph detection is strong, with 88.8% of PARA_BREAK tokens correctly identified.

Our dual-head model achieves 87.1% recall on REMOVE tokens—a reduction in the miss rate from 63.0% to 12.9%. It correctly identifies 99.3% of KEEP tokens with only 0.7% false-positive REMOVE predictions, and maintains 88.8% paragraph detection accuracy.

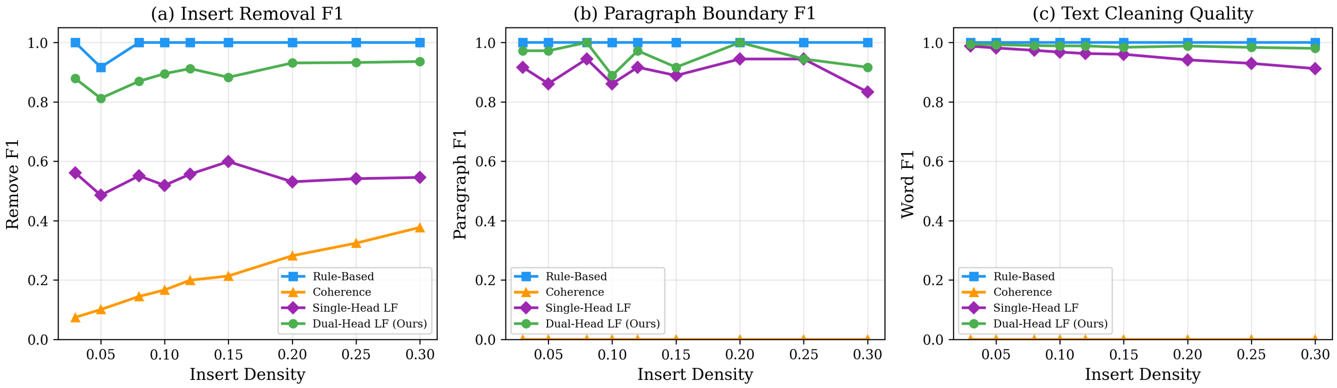

Figure 3: Performance across insert density levels (3–30%). (a) REMOVE-class F1: our dual-head model maintains 0.81–0.94 while the single-head baseline plateaus at 0.49–0.60. (b) Paragraph boundary F1: both Longformer variants achieve strong segmentation. (c) Word-level cleaning quality: our model stays above 0.98 across all densities.

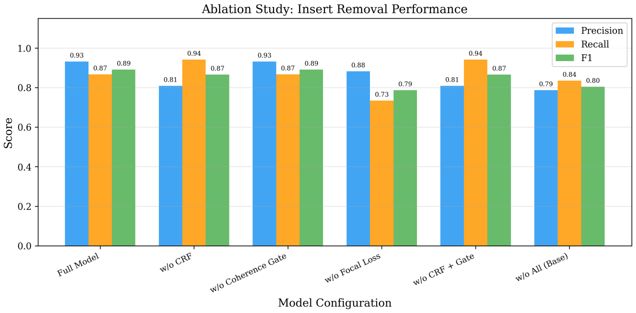

Figure 4: Ablation study: precision, recall, and F1 for insert removal under different model configurations. Focal loss provides the largest individual contribution to F1.

3.5 Data Characteristics

Figure 6 characterizes the synthetic dataset. The label distribution shifts predictably with insert density: at 5% density, KEEP tokens comprise 92.7% and REMOVE tokens 5.3%, while at 25% density these shift to 78.9% and 19.4% respectively. PARA_BREAK tokens remain constant at approximately 1.7–2.0% across all densities.

Insert spans have a mean length of 3.0 tokens (std: 1.7) at 15% density, with a right-skewed distribution ranging from 1 to 8 tokens. This variability motivates the CRF layer, which enforces minimum span constraints to reduce isolated single-token false positives.

4 CONCLUSION

We addressed the open problem identified by Bondarenko et al. [4] where a single-head Longformer model failed to remove out-of-context inserts from dictation transcripts despite successfully segmenting paragraphs. Our dual-head architecture decomposes the task into two specialized sub-tasks—paragraph segmentation and insert detection—with a coherence gating mechanism and CRF fusion layer.

The key finding is that the architectural decomposition itself provides the largest improvement: the base dual-head model without CRF, gate, or focal loss already achieves 0.805 REMOVE F1 compared to 0.514 for the single-head baseline. Adding focal loss provides the next largest gain (+0.086), addressing the fundamental class imbalance where KEEP tokens comprise 85–93% of the data. The CRF layer contributes a further +0.025 by improving precision through span-level constraints.

Our approach maintains strong paragraph segmentation (F1 = 0.908) and high overall text cleaning quality (word F1 = 0.988), demonstrating that the insert removal improvement does not come at the expense of other sub-tasks. The model achieves consistent performance across insert densities from 3% to 30%, with REMOVE F1 ranging from 0.812 to 0.936.

Future work should address three limitations: (1) evaluation on real dictation transcripts rather than synthetic data, (2) extension to multilingual settings where dictation commands may be in a different language than the literary text, and (3) integration with end-to-end ASR systems for joint optimization. The synthetic data pipeline and dual-head architecture provide a foundation for addressing the broader challenge of structured dictation transcript cleaning.

REFERENCES

[1] Nguyen Bach and Fei Huang. 2019. Noisy BiLSTM-Based Models for Disfluency Detection. In *Proc. Interspeech*.

[2] Regina Barzilay and Mirella Lapata. 2008. Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics* 34, 1 (2008), 1–34.

[3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150* (2020).

[4] Ilia Bondarenko et al. 2026. Pisets: A Robust Speech Recognition System for Lectures and Interviews. *arXiv preprint arXiv:2601.18415* (2026).

[5] Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752* (2023).

[6] Paria Jamshid Lou and Mark Johnson. 2020. Improving Disfluency Detection by Self-Training a Self-Attentive Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3754–3763.

[7] John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. 282–289.

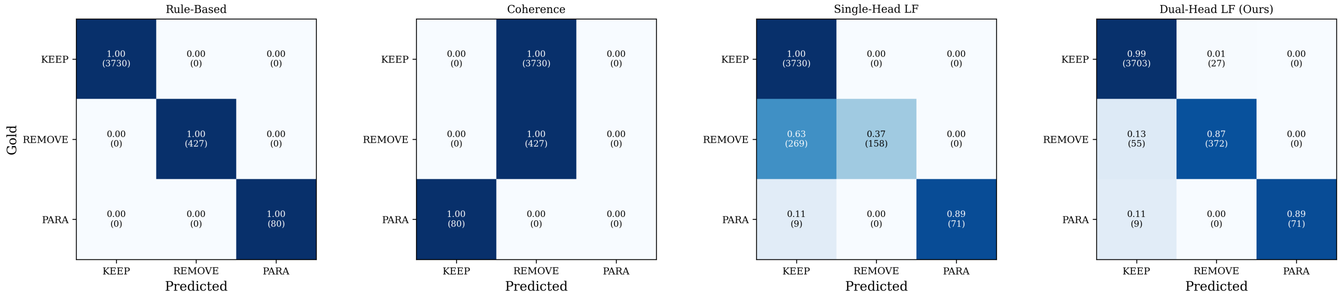

Figure 5: Normalized confusion matrices for all four methods. The single-head Longformer misclassifies 63.0% of REMOVE tokens as KEEP, while our dual-head model reduces this to 12.9%.

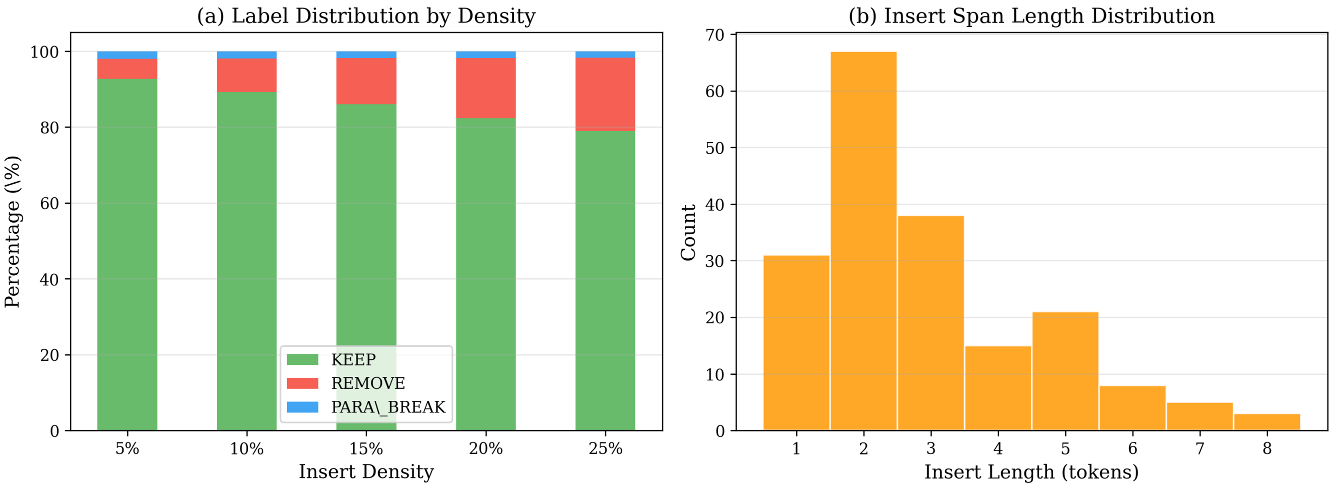

Figure 6: Dataset characteristics. (a) Label distribution by insert density: REMOVE tokens range from 5.3% at low density to 19.4% at very high density. (b) Insert span length distribution at 15% density, showing a mean length of 3.0 tokens.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2980–2988.

[9] Shaolei Wang, Wanxiang Che, and Ting Liu. 2020. Multi-Task Self-Training for Disfluency Detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 5765–5775.

[10] Jinan Xu, Shuming Li, and Deyi Li. 2019. Cross-Lingual Transfer Learning for Text Coherence Assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 5 (2019), 930–941.

[11] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems*, Vol. 33.

[12] Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency Detection Using a Bidirectional LSTM. In *Proc. Interspeech*.