Team Aurum at MedExACT 2026@ACL: Data Augmentation and Clinical Longformer Fine-Tuning for Medical Decision Extraction

Vinay Babu Ulli
Oogwai Analytics,
Bangalore, India
ullivinaybabu@gmail.com

Jyoti Kumari
Department of Linguistics,
Banaras Hindu University, India
jyoti@bhu.ac.in

Anindita Mondal
Language Technologies
Research Center,
IIIT Hydearabad, India
anindita.mondal@research.iiit.ac.in

Abstract

This paper describes the system submitted by team **Aurum** to the Medical Decision Extraction, Analysis, and Classification Task (MedExACT) at BioNLP 2026. The task requires the extraction and classification of contiguous text spans representing medical decisions from lengthy ICU discharge summaries. To address the dual challenges of long document lengths and severe class imbalance within a limited training set of 350 notes, we propose a two-pronged strategy. First, we employ a tripartite data augmentation pipeline utilizing rule-based entity replacement, LLM-based contextual paraphrasing, and synthetic note generation to expand the training data to over 2,300 notes. Second, we fine-tune a domain-specific Clinical Longformer model equipped with a sliding-window inference mechanism and Focal Loss to handle sequences up to 2,048 tokens while focusing on rare decision categories. Paired with a targeted post-processing module, our system achieved a Final Score of 0.5251, demonstrating high token-level detection (Token F1: 0.6311) and strong stability across patient demographics.

1 Introduction

Clinical discharge summaries document critical medical decisions, ranging from diagnostic assessments to therapeutic procedures, that shape downstream patient care. However, these decisions are embedded in dense, unstructured, and often lengthy free-text narratives (Agrawal et al., 2022). The MedExACT 2026 shared task (Elgaar et al., 2026) aims to advance the automatic extraction of these decisions by framing it as a joint span detection and classification problem. Crucially, systems are evaluated not only on average performance but also on their worst-group robustness across demographic lines.

Extracting medical decisions from the MedDec dataset (Elgaar et al., 2024) presents two primary bottlenecks. First, the training set is relatively small (350 notes), leading to rapid model overfitting. Second, discharge summaries average between 2,000 and 6,000 tokens in length. Standard Transformer encoders limited to 512 tokens (e.g., ClinicalBERT, ELECTRA) inevitably truncate 60–80% of the text, discarding vital sections like the *Hospital Course*.

To overcome these limitations, we engineered a pipeline that combines extensive **Data Augmentation** with **Clinical Longformer Fine-Tuning**. Our augmentation strategies expand linguistic and entity diversity, while the Longformer architecture (Beltagy et al., 2020) ensures that decisions located deep within the document are successfully processed.

2 Task Description and Background

The MedExACT 2026 task is built upon the MedDec dataset (Elgaar et al., 2024), derived from the MIMIC-III critical care corpus (Johnson et al., 2016). The utility of this dataset extends beyond static benchmarks, having recently supported the development of MedDecXtract, an interactive downstream tool designed to assist clinicians in extracting, visualizing, and annotating medical decisions in real-time (Elgaar et al., 2025).

Task Definition: Given a full discharge summary, systems must detect contiguous text spans that express medical decisions and assign each span one of nine DICTUM (Ofstad et al., 2016) decision categories (*Contact related, Gathering information, Defining problem, Treatment goal, Drug, Therapeutic procedure, Evaluating test result, Deferment, Advice and precaution*) or a *None* label when no decision is present.

Note: Some MedDec annotations may include Category 10 (Legal/insurance related) and Category 11 (Others). These categories are out of scope for MedExACT@ACL 2026 and are ignored by the official evaluator.

Proceedings of the BioNLP 2026 (Shared Tasks), pages 224–228

Decision Type	Sex		Race						Lng. Proficiency	
(lr)2-3 (lr)4-9 (lr)10-11	**Male**	**Female**	**White**	**AA**	**Hispanic**	**Asian**	**NH**	**Other**	**En**	**Non-En**
	(n=259)	(n=192)	(n=327)	(n=42)	(n=25)	(n=15)	(n=1)	(n=21)	(n=260)	(n=45)
Defining Problem	39.2	38.8	39.5	37.5	38.0	36.4	30.9	38.6	38.7	39.2
Drug	26.0	25.1	25.7	24.4	25.0	27.5	19.1	27.0	26.1	25.6
Evaluation	12.9	13.6	12.6	16.6	13.3	12.7	25.5	12.8	13.1	13.9
Therapeutic proc.	12.2	12.4	12.4	12.5	11.7	13.2	10.6	12.2	12.0	12.0
Contact	4.9	5.2	5.0	4.6	6.0	5.4	8.5	4.3	4.8	5.1
Advice	3.4	3.5	3.5	3.2	4.2	3.3	0.0	3.9	3.9	3.0
Gathering info	0.8	0.9	0.8	0.7	1.2	1.3	5.3	0.9	0.9	0.6
Treatment goal	0.3	0.3	0.3	0.3	0.4	0.2	0.0	0.2	0.2	0.4
Deferment	0.2	0.2	0.2	0.2	0.2	0.0	0.0	0.1	0.2	0.2
Legal/Insurance	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total Count	33,054	24,235	41,666	5,684	3,264	1,737	94	3,078	37,026	6,295

Table 1: Percentage of annotated spans for each decision category across protected variables in the MedDec dataset. *n* is the number of discharge summaries for each category and the last row shows the total count of decisions per variable.

Dataset and Demographics: The dataset exhibits extreme class imbalance. As shown in Table 1, *Defining Problem* and *Drug* decisions dominate the annotations, accounting for roughly 65% of all spans. Conversely, categories such as *Treatment Goal*, *Gathering info*, and *Deferment* each represent less than 1% of the data. Table 1 details the distribution of these spans across protected variables, which is fundamental to the task's worst-group robustness evaluation.

3 Methodology

We formulate the medical decision extraction task as a token-level sequence tagging problem. We apply a standard BIO (Begin, Inside, Outside) labeling scheme to the 9 in-scope DICTUM categories, resulting in 19 distinct prediction classes (9 categories $\times$ 2 + 1 'O'-tag). Our methodology comprises a multi-strategy data augmentation pipeline, a domain-adapted Clinical Longformer backbone, specialized loss functions, and targeted post-processing.

3.1 Tripartite Data Augmentation Pipeline

With only 350 training notes and models containing over 140 million parameters, overfitting and memorization are significant risks. To prevent the model from simply memorizing the training set, we designed a three-pronged augmentation strategy to inject entity, linguistic, and volume diversity, thereby expanding the training corpus to over 2,300 notes.

Strategy 1: Rule-Based Entity Replacement. To increase entity diversity, we utilized curated medical dictionaries to replace specific drugs, conditions, and therapeutic procedures with clinically similar alternatives (e.g., swapping *vancomycin* with *meropenem*, or *pneumonia* with *bronchitis*). Replacements were executed with a 30% probability per entity. Crucially, because string lengths vary, our algorithm dynamically recalculates and shifts the exact character offsets for all subsequent annotations in the document. This zero-cost method generated roughly 1,050 augmented notes.

Strategy 2: LLM Contextual Paraphrasing. To increase linguistic diversity and make the model robust to different physician writing styles, we leveraged the Claude Sonnet Large Language Model (LLM) via the OpenRouter API. We parsed the training documents into decision spans and the non-annotated "gaps" between them. We prompted the LLM to rewrite only the gap texts while keeping the annotated spans strictly verbatim. This forces the model to learn the semantic boundaries of decisions regardless of surrounding syntax (e.g., altering "*was diagnosed with [SPAN]*" to "*workup was consistent with [SPAN]*"). This process yielded approximately 700 paraphrased notes.

Strategy 3: Synthetic Note Generation. Categories such as *Treatment Goal* (Cat 4), *Gathering Info* (Cat 2), and *Deferment* (Cat 8) are severely starved, appearing in less than 1% of annotations. We prompted the LLM to generate entirely synthetic discharge summaries targeting these rare categories using the following prompt template:

> *"Generate a realistic hospital discharge summary for a patient with [Disease]. Include 40-80 medical decisions across categories 1-9. Return the note text AND a JSON list of annotations with exact character offsets."*

For the sampling strategy, we utilized a temperature of 0.7 and Top-P of 0.9 to ensure varied but clinically plausible outputs. Because LLM-generated character offsets are notoriously inaccurate (often hallucinating string lengths), we implemented strict programmatic filtering: the system searched the generated text for the exact predicted string. Annotations that could not be perfectly substring-aligned were discarded. This targeted generation successfully increased the representation of starved categories (Cat 2, 4, 8) from under 1% to approximately 6% of the total training spans, yielding 200 high-quality synthetic notes.

Augmentation Quality and Validation. To validate the quality of our augmented data, we manually inspected a random sample of 50 notes from each strategy. Entity replacement maintained near 100% clinical plausibility as it relied strictly on curated medical dictionaries (e.g., matching drug classes). LLM paraphrasing preserved the exact decision spans, introducing minor syntactic artifacts in only $\sim$4% of the surrounding gaps. For the synthetic generation, manual inspection revealed an initial LLM offset error rate of roughly 20–25%; however, our strict substring-matching filter effectively eliminated these misalignments, dropping the final annotation error rate in the synthetic set to under 2% while maintaining coherent clinical narratives.

3.2 Clinical Longformer Architecture

Discharge summaries in the MedDec dataset average between 2,000 and 6,000 tokens. Standard Transformer encoders (e.g., BERT, ELECTRA) impose a 512-token limit, which effectively truncates 60% to 80% of a standard discharge summary, completely discarding critical sections such as the *Hospital Course* where the majority of medical decisions reside.

To resolve this, we utilized the Longformer architecture (Beltagy et al., 2020), which replaces $\mathcal{O}(n^2)$ full self-attention with a sparse attention pattern combining a sliding local window (512 tokens on each side) and task-specific global attention on the `[CLS]` token. This reduces complexity to $\mathcal{O}(n \times w)$ and extends the maximum sequence length to 4,096 tokens.

Specifically, we selected `yikuan8/Clinical-Longformer`, which was further pre-trained on MIMIC-III clinical notes. This domain-specific pre-training ensures the model tokenizer is adapted to medical vocabulary and inherently understands discharge summary structures (e.g., section headers, de-identification markers).

Layer Freezing and Custom Classification Head. To prevent catastrophic forgetting of pre-trained domain knowledge and to stabilize training on our small dataset, we froze the bottom 6 layers of the 12-layer encoder. This reduced the trainable parameter count from 149M to approximately 75M. The output of the trainable top 6 layers was fed into a custom 2-layer classification head: a linear projection from 768 to 256 dimensions, followed by a GELU activation, a 0.2 dropout layer, and a final projection to the 19 BIO classes.

3.3 Addressing Class Imbalance with Focal Loss

The class distribution in the MedDec dataset is exceptionally skewed, with the ‘O‘ (Outside) tag comprising 95.2% of all tokens in the training data. Under standard Cross-Entropy loss, the model easily achieves high overall accuracy by defaulting to ‘O‘ predictions, thereby ignoring the sparse medical decision boundaries.

To penalize the model for relying on easy background tokens, we replaced Cross-Entropy with Focal Loss (Lin et al., 2017):

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \qquad (1)$$

We set the focusing parameter $\gamma = 2.0$ to aggressively down-weight easily classified examples. Furthermore, we applied a class-specific weighting factor $\alpha_O = 0.3$ strictly to the ‘O‘ tag, ensuring that the loss gradients are dominated by the hard, rare decision boundaries (Categories 1–9) rather than background medical text.

3.4 Inference Strategy and Post-Processing

Due to GPU memory constraints during training, sequences were cropped to a maximum length of 2,048 tokens. However, during evaluation, we implemented a non-overlapping sliding window inference mechanism. The document is processed in 2,048-token chunks, and the resulting hidden states are concatenated before being passed to the classification head. This ensures that a single, unified prediction array is generated for notes of any length.

Converting token-level predictions back to the exact character-level offsets required by the official shared task evaluator introduces further

challenges. Standard forward-searching token-to-character mapping frequently misaligns boundaries. We resolved this by implementing a bidirectional search mapping, which eliminated thousands of near-miss offset errors.

Finally, error analysis on our validation set highlighted systematic failure modes that we addressed via a targeted rule-based post-processing pipeline:

1. **Boundary Cleaning:** Striping trailing punctuation (e.g., periods, commas) and whitespaces from predicted spans to match the strict evaluator.
2. **Regex Recovery for Rare Classes:** Employing regular expressions to identify and tag highly specific, heavily missed patterns for *Deferment* (e.g., "*deferred to outpatient follow-up*") and *Gathering Info* (e.g., "*ordered CT of the chest*").
3. **Deduplication:** Removing completely identical spans that occasionally arise at the boundaries of sliding inference windows.

4 Experimental Setup

Models were trained using PyTorch and HuggingFace. We trained the Clinical Longformer for 5,000 steps using the AdamW optimizer with a learning rate of 4e-5 and a 10% linear warmup. Due to the memory footprint of 2,048-token sequences, we utilized a batch size of 4 with 2 gradient accumulation steps (effective batch size of 8). Checkpoints were evaluated every 500 steps, and the best model was selected based on validation Span F1.

5 Results and Analysis

Our system, submitted under the team name *Aurum*, achieved the results detailed in Table 2 on the official hidden test set.

Metric	Score
Final Score	**0.5251**
Base Score	0.5362
Worst Group Score	0.5140
Token F1	0.6311
Span F1	0.4414

Table 2: Official hidden test set results for Team Aurum.

5.1 Ablation Study

To isolate the contribution of each system component, we conducted an ablation study on our local validation split. Table 3 illustrates the incremental performance gains starting from a standard 512-token baseline.

Model Configuration	Token F1	Span F1
Baseline (ELECTRA-Base, 512)	0.452	0.301
+ Clinical Longformer (2048)	0.584	0.387
+ Data Augmentation	0.612	0.415
+ Focal Loss	0.625	0.430
+ Post-Processing (Final)	**0.631**	**0.441**

Table 3: Incremental ablation study showing the impact of each architectural and methodological choice.

The transition from a 512-token baseline to the Clinical Longformer yielded the most substantial improvement (+0.132 Token F1), confirming that document truncation was the primary bottleneck. The tripartite data augmentation pipeline provided the second-largest boost (+0.028 Token F1), validating our hypothesis that injecting linguistic and rare-class diversity mitigates overfitting. Finally, Focal Loss and rule-based post-processing provided targeted gains, specifically improving boundary alignment and the recall of heavily starved classes like *Deferment*.

5.2 Error Analysis

Span vs. Token F1 Trade-off: The results highlight a pronounced gap between Token F1 (0.6311) and Span F1 (0.4414). This pattern aligns with the task organizers' observations regarding the difficulty of exact span extraction. Our high Token F1 indicates that the Longformer successfully locates the correct semantic regions; however, the strict character-level exact-match required by the Span F1 evaluator harshly penalizes minor boundary discrepancies (e.g., including an extra article like "a" or missing a trailing word). Precise span calibration remains a bottleneck.

Subgroup Robustness: A key feature of our performance is the minimal gap between the Base Score (0.5362) and the Worst Group Score (0.5140), a difference of only 0.022 points. This indicates that our model exhibits strong fairness across the evaluated demographic buckets. We attribute this robustness to two factors: the extensive linguistic variety introduced by our LLM paraphrasing augmentation, and the inherent demographic representation captured within the MIMIC-III pretraining of the Clinical Longformer backbone.

6 Conclusion

In this paper, we presented team *Aurum*'s submission to the MedExACT 2026 shared task. By replacing standard 512-token encoders with a Clinical Longformer, we solved the critical issue of document truncation in ICU discharge summaries. Furthermore, leveraging a multi-strategy data augmentation pipeline and Focal Loss allowed us to successfully train a robust model on a small, highly imbalanced dataset. While our approach yielded strong demographic stability and token-level detection, future work must focus on advanced boundary-refinement modules to bridge the gap between Token F1 and exact-match Span F1.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Mohamed Elgaar, Hadi Amiri, Mitra Mohtarami, and Leo Anthony Celi. 2025. MedDecXtract: A clinician-support system for extracting, visualizing, and annotating medical decisions in clinical narratives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 481–489, Vienna, Austria. Association for Computational Linguistics.

Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. MedDec: A dataset for extracting medical decisions from discharge summaries. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455, Bangkok, Thailand. Association for Computational Linguistics.

Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Mehrnaz Sadrolashrafi, Mitra Mohtarami, Adrian Wong, Hadi Amiri, and Leo A. Celi. 2026. Overview of medical decision extraction, analysis, and classification task (medexact) 2026. In *The 25th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, San Diego, California, USA. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1).

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*, pages 2980–2988.

Eirik H Ofstad, Jan C Frich, Edvin Schei, Richard M Frankel, and Pål Gulbrandsen. 2016. What is a medical decision? a taxonomy based on physician statements in hospital encounters: a qualitative study. *BMJ Open*, 6(2):e010098.