CanSA at MedExACT@ACL 2026: Zero-Shot, Fine-Tuned, and Retrieval-Augmented Extraction of Clinical Decisions with Corpus Boundary Diagnostics

Mohammed Alliheedi[1]*, Robert E. Mercer[2]*, Anemily Vincens Machina[2]*, Sudipta Singha Roy[2]*, Yetian Wang[3]*, Xindi Wang[4]*

[1]Al-Baha University, Saudi Arabia
[2]University of Western Ontario, Canada
[3]University of Waterloo, Canada
[4]Shandong University, China
*These authors contributed equally to this work.

malliheedi@bu.edu.sa, {rmercer,anemily.machina, ssinghar}@uwo.ca,
yetian.wang@uwaterloo.ca, xindi.wang@sdu.edu.cn

Abstract

We present the CanSA system for the MedExACT@ACL 2026 shared task, which requires extracting and classifying clinical decisions from ICU discharge summaries into nine DICTUM categories. We have developed three approaches: (1) a training-free system which consists of a preprocessing module that normalizes text and an inference engine combining zero shot LLMs with a RAG ensemble, (2) a supervised fine-tuning method which required training, and (3) a training-free retrieval-augmented pipeline employing TF–IDF-based lexical retrieval to surface in-context exemplars from the development corpus, combined with section-aware chunking and structured extraction calls to a large language model. Our team's best submission achieved a Final Score of 0.41, ranking 34th out of 37 on the official test leaderboard.

1 Introduction

The MedExACT shared task at the BioNLP 2026 workshop (Elgaar et al., 2026) challenges systems to extract exact text spans representing medical decisions from Intensive Care Unit (ICU) discharge summaries and classify them according to the DICTUM taxonomy (Ofstad et al., 2016; Elgaar et al., 2024). Because the evaluation heavily penalizes boundary misalignment, systems must demonstrate both deep semantic classification and strict character level precision.

Team CanSA presents three methodologies. The first addressed this challenge by building a modular, training-free inference pipeline. To avoid the heavy costs of task specific fine tuning, we evaluated several leading models, specifically Llama 3, Qwen, and Kimi, by deploying them through local Ollama instances on an NVIDIA A6000 and remote NVIDIA NIM API endpoints. The second provided a supervised fine-tuning of a bi-directional encoder model on a token labelling task. And lastly, employed a training-free, retrieval-augmented pipeline in which TF-IDF-based lexical retrieval surfaces in-context exemplars from the development corpus, and a large language model performs sequential span proposal and structured extraction over section-aware chunks of each discharge note, followed by schema-based validation and automated repair. We also contribute a rigorous error analysis of the MedDec corpus by auditing 104 same category overlapping annotation pairs providing actionable insights for future clinical dataset curation.

The remainder of this paper is organized as follows: Section 2 details the specifics of each of the three methodologies. Section 3 outlines our experimental setup. Section 3.2 presents the shared task results and post deadline experiments, followed by the corpus audit in Section 4. We conclude in Section 5.

2 System Architectures

2.1 Training-Free Inference Pipeline

The MedExACT task requires exact character offset matching. During initial pipeline development, we observed severe boundary extension errors caused by standard tokenizers splitting at MIMIC III de identification markers (e.g., `[**...**]`). To mitigate this, we implemented a strict length preserving normalization heuristic: replacing all newline characters (`\n`) with spaces. This single transformation resolved tens of thousands of offset mis-

Proceedings of the BioNLP 2026 (Shared Tasks), pages 141–145
July 3-4, 2026

matches across the raw clinical text.

Following normalization, documents are segmented using a section aware sliding window. To optimize inference efficiency and reduce API costs, the pipeline explicitly skips the *Social History* and *Family History* sections, as corpus analysis revealed zero gold standard annotations in these segments. Finally, a post processing filter enforces a minimum annotation length of five characters to eliminate single token model artifacts.

Our primary baseline module relies on zero shot extraction using frontier LLMs. The system prompts the model with a strict JSON schema definition encompassing the nine DICTUM categories.

A critical engineering challenge involved managing models equipped with implicit reasoning pathways (e.g., Qwen and DeepSeek). These models frequently output reasoning chains (often enclosed in `<think>` tags) that corrupt the required JSON structure. To enforce deterministic formatting, we implemented a dual suppression strategy: prepending a `/no_think` directive to the user prompt and explicitly setting `enable_thinking: false` in the API payload.

To evaluate the impact of in context learning without task specific fine tuning, we developed an alternative RAG enhanced pipeline. We indexed the MedDec training dataset using FAISS GPU (Johnson et al., 2019). For each sliding window chunk of the target document, the system retrieves the top k most semantically similar training examples and appends them to the prompt as few shot exemplars, attempting to ground the model's predictions in gold standard phrasing.

Initial development was conducted locally on an NVIDIA A6000 GPU (48 GB VRAM) using Ollama to host a high density ensemble. However, hardware constraints necessitated a transition to a constrained local RTX 3060 (12 GB VRAM) setup, relying on NVIDIA NIM API endpoints to access models exceeding 70 billion parameters.

This remote transition introduced severe operational bottlenecks. We observed that the increased token payload of the RAG pipeline highly correlated with HTTP 504 (Gateway Timeout) and HTTP 429 (Too Many Requests) errors. These connection drops frequently caused truncated JSON responses mid transit—a degradation effect that forced zero span extractions for the affected chunks. Consequently, the final pipeline enforces a stability first configuration, utilizing a strict 15 second API delay and exponential backoff to ensure continuous execution during the evaluation phase.

2.2 Supervised Fine-tuning

The supervised learning method fine tuned a bi-directional encoder model: BAAI/bge-m3 (Chen et al., 2023). The goal was to start with a single label classification task, and then use the learned classifiers to extend to multi-label classification. A bi-directional encoder was used, compared to a auto-regressive Language Model, as the extra right side context should be important for labelling tokens and spans. To create a single label for each token, any token with multiple labels was assigned the most frequent label as they appear in the training dataset. Unfortunately, time did not permit the multi-label classification, so only results for the single label task are reported.

The model was trained using Cross Entropy Loss weighted so that each class had equal contribution. The Huggingface training interface was used: it is relatively straightforward to use and hack as needed; it allows multiple GPUs with no extra effort. Parameters used: learning rate 2e-4, per device batch size of 1 with gradient accumulation of 16 steps, 100 training epochs with best checkpoint per epoch retrained based on validation F1 score, 100 total epochs with a warmup of 110 steps. Three different seeds were used after which final token labels on the test set were based on the soft ensemble, summation followed by argmax, of all three best checkpoints.

Given more time, other foundation models would be used for fine-tuning on the single label task, and the classification vectors from those training runs would be used to seed a binary classifier for each label: with bias initialization tuned on the training set.

2.3 Training-Free Retrieval-Augmented Pipeline

Span-level adverse-event–style extraction from long discharge summaries is addressed through a training-free, retrieval-augmented language model pipeline. Each discharge note is segmented into clinical sections. Following task conventions, predicted spans overlapping allergy regions are removed in a post-hoc filtering step. The remaining text is split into section-aware chunks of at most 5,000 characters, ensuring that each language model call fits within context limits while coarse document structure is preserved.

A lexical index is constructed over the development corpus: for every document with paired raw text and reference annotations, a truncated prefix of the discharge (up to 80,000 characters per note) is indexed together with a compact serialization of up to 40 gold spans comprising shortened decision text and category labels. Documents are represented using TF–IDF features (unigrams and bigrams; maximum 50,000 features; document-frequency cap of 0.95). At inference time, each chunk's text serves as the query; training documents are ranked by cosine similarity and the top three matches are retrieved, excluding the current discharge ID to prevent trivial self-retrieval. For each retrieved document, a truncated text excerpt (up to 3,500 characters) along with the corresponding compact gold annotations is incorporated into the prompt as in-context exemplars.

Claude Sonnet 4 is employed at temperature zero. For each chunk, two sequential calls are issued: (1) a span proposal step, conditioned on the chunk text, section headers, and the retrieval block; and (2) a structuring step, in which the proposals are converted into JSON with chunk-local character offsets, span text (referred to as *decision*), and a category label. Local offsets are subsequently shifted to global UTF-8 offsets within the full note. Where a predicted span text does not exactly match the underlying substring, local realignment is attempted within a window around the chunk position, including whitespace-tolerant matching. Annotations from all chunks are then merged, deduplicated by start position, end position, and category, sorted, and assigned stable identifiers; spans overlapping allergy regions are discarded.

The assembled document-level JSON is validated against a schema tied to the full raw text. In cases where validation fails, a single additional language model call is issued, conditioned on the entire discharge note and the current prediction; the model is instructed to rewrite the full annotation set so as to satisfy all constraints. Allergy-region filtering is subsequently re-applied where applicable. Finally, category labels are mapped to the numeric codes required by the shared-task evaluation format.

3 Experimental Setup and Results

3.1 Evaluation Metrics

System performance is evaluated using the official MedExACT Final Score, defined as the arithmetic mean of the Base Score and the Worst Group Score (a fairness metric evaluated across patient demographic subgroups). The Base Score itself is the arithmetic mean of Span F1 (requiring exact character boundary matches) and Token F1 (measuring word level overlap). Throughout our diagnostics, we track the differential between Token F1 and Span F1 to isolate boundary tokenization artifacts from clinical reasoning errors.

3.2 Official Shared Task Results

For the official MedExACT test phase, Team CanSA submitted three independent runs to the Codabench leaderboard. Our training free zero shot pipeline using Qwen 3.5 achieved a Span F1 of 0.31 and a Token F1 of 0.44 (resulting in a Base Score of 0.38) alongside a Final Score of 0.34.

For the supervised fine-tuning method, the best checkpoints had training/eval token F1 scores 0.93/0.63, 0.87/0.61, and 0.91/0.62. Note: this is on the single label task. On the validation set, the ensemble labels achieved a 0.61 validation and span annotation F1 of 0.16 (task score 0.38) with sub group scores from 0.28 to 0.43. The final score on the validation set was 0.33. The leaderboard score for this method was 0.37.

The training-free retrieval-augmented pipeline achieves a span F1 of 0.35, token F1 of 0.51, and a final score of 0.38 on the validation set. On the official test set, a span F1 of 0.34, token F1 of 0.50, and a final leaderboard score of 0.41 are obtained.

Our highest performing submission achieved a Base Score of 0.42 and a Final Score of 0.41, placing 34th out of 37 on the official test set leaderboard.

3.3 Post Deadline Architectural Analysis

To properly evaluate the architectural trade offs of the first pipeline, we conducted post deadline experiments on the complete validation set. Table 1 summarizes the performance of frontier models, scored strictly via the official MedExACT evaluation script. An analysis of these results highlights two critical operational findings:

The RAG API Penalty. In a high latency remote environment, appending FAISS retrieved context dynamically increases payload size, correlating strongly with HTTP 504 timeouts and truncated JSON responses. While models like Kimi K2 proved highly resilient, other models suffered from mid transit connection drops degrading overall throughput.

Configuration	Span F1	Tok F1	Base	Final
Kimi K2 (Zero-Shot)	0.28	0.43	0.35	0.32
Kimi K2 (RAG Ens.)	0.29	0.44	0.36	**0.34**
Qwen 3.5 (Zero-Shot)	**0.31**	**0.44**	**0.38**	**0.34**
Qwen 3.5 (RAG Ens.)	0.28	0.43	0.35	0.30
Llama 3.3 (Zero-Shot)	0.29	0.42	0.36	0.31
Mistral (Zero-Shot)	0.27	0.43	0.35	0.30

Table 1: Validation set ($n = 53$) performance across NVIDIA NIM API endpoints.

The Token to Span Gap. Across all evaluated models, Token F1 consistently outperformed Span F1 by a margin of 0.13 to 0.16. For example, while Mistral achieved a peak Token F1 of 0.43, its strict Span F1 dropped to 0.27. This massive differential proves that while frontier LLMs possess the semantic capability to identify clinical decisions, they fail to predict the exact character boundaries required by the MedDec gold standard.

4 Corpus Audit: The Boundary Gap

Category	Tokens	Gap Size	Count
1	12.1 ± 25.7	178.5 ± 1959.3	6.5 ± 9.7
2	8.7 ± 10.9	276.1 ± 4400.4	1.1 ± 4.9
3	6.9 ± 69.3	23.2 ± 64.1	24.6 ± 175.4
4	6.4 ± 3.4	294.5 ± 5023.8	0.3 ± 3.2
5	8.1 ± 55.5	31.9 ± 124.7	31.2 ± 85.1
6	5.7 ± 39.9	79.6 ± 445.4	15.2 ± 20.1
7	12.8 ± 43.8	44.7 ± 184.9	16.6 ± 28.2
8	9.3 ± 14.8	358.1 ± 6756.6	0.24 ± 3.08
9	19.6 ± 27.5	36.5 ± 197.9	4.3 ± 7.2

Table 2: Annotation statistics at the document level: Tokens (annotation size); Gap Size (tokens between annotations); Count (annotations in document).

The persistent 0.13 to 0.16 differential between Token F1 and Span F1 highlighted in Section 3.3 indicates that while frontier LLMs successfully identify clinical concepts, they don't always map these concepts to the exact character boundaries defined by the gold standard. To investigate the structural cause of this gap, we conducted a comprehensive annotation quality audit of the MedDec corpus and a token level analysis of annotation statistics (see Table 2).

We identified 104 same category overlapping annotation pairs distributed across 79 documents (representing roughly 17.5% of the validation corpus). A manual review of these pairs revealed three distinct fragmentation typologies:

- **Boundary Extension Errors (n=21):** These are artificial splits caused by tokenizer artifacts, specifically where automated tools (e.g., Stanza) over extended annotation boundaries adjacent to MIMIC III de-identification markers (e.g., `[**...**]`).
- **Near Identical Duplicates (n=7):** Annotations covering the exact same semantic clinical concept with minor 1 to 2 character shifts.
- **Genuine Subspans (n=74):** Hierarchical or nested clinical decisions, of which 33 were fragment level annotations under 20 characters in length.

The presence of these structural artifacts, particularly the 21 boundary extension errors linked to de-identification masking, imposes a hard recall ceiling on exact match extraction systems. Because strict span evaluation harshly penalizes partial boundary misalignment, heuristic post processing tools cannot systematically recover these points without introducing false positives. Consequently, we argue that the Token to Span gap is largely a reflection of corpus level tokenization methodology rather than a deficit in LLM clinical reasoning.

5 Conclusion

In this paper, we described the CanSA system for the MedExACT@ACL 2026 shared task. Our team developed three approaches including a training-free extraction pipeline that combines length preserving text normalization, stable API management, and zero shot LLM inference, a supervised fine-tuning approach, and a training-free retrieval-augmented pipeline employing TF–IDF-based lexical retrieval to surface in-context exemplars from the development corpus combined with section-aware chunking, achieving a Final Score of 0.41 on the official test leaderboard (Rank 34/37). Through extensive post deadline evaluation on the first method, we demonstrated the "RAG API Penalty" where increased payload sizes trigger silent mid transit connection timeouts and highlighted a systemic Token to Span F1 gap. Our corpus audit links this boundary fragmentation directly to tokenizer artifacts at de-identification markers. We conclude that while frontier LLMs possess strong clinical semantic capabilities, their deployment in exact match extraction tasks requires stability-first engineering and necessitates tokenizer aware corpus curation for future clinical benchmarks.

Ethics and Data Statement

This work uses the MIMIC III critical care database (Johnson et al., 2016), which is a restricted dataset requiring credentialed access via PhysioNet. All analyses were conducted in compliance with the MIMIC III data use agreement. No patient re identification was attempted. The MedDec annotations (Elgaar et al., 2024) are publicly available for research purposes.

References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2309.07597.

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, Li-wei 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):1–9.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Eirik H Ofstad, Jan C Frich, Edvin Schei, Richard M Frankel, and Pål Gulbrandsen. 2016. Dictum: a taxonomy of medical decisions in clinical encounters. *BMJ open*, 6(2):e010078.