CASPAR: A Context-Aware Span Refinement Approach for Decision Support

Jing Tao
School of Computing
Queen's University
Kingston, ON, Canada
jing.tao@queensu.ca

Amir Eskandari
School of Computing
Queen's University
Kingston, ON, Canada
Amir.Eskandari@queensu.ca

Farhana Zulkernine
School of Computing
Queen's University
Kingston, ON, Canada
farhana.zulkernine@queensu.ca

Abstract

Extracting structured clinical information by selecting a text span from long medical discharge summaries using Natural Language Processing (NLP) poses difficult challenges. Heterogeneity of the text, semantic overlap of the decision categories, and uncertainty in identifying boundaries of text span aligned with highly unbalanced labels. Contribute to these challenges, the MedExACT competition in BioNLP @ ACL 2026 formulates the medical concept and text span extraction problem from unstructured text as a classification task over nine DICTUM decision categories. Unstructured ICU discharge summaries are used from the MedDec dataset, for solution approach across eight demographic subgroups. We propose a two-stage data processing approach for the text span extraction and classification task where Stage 1 performs document-level sequence labeling with a pipeline composed of a RoBERTa-base, a Bidirectional Gated Recurrent Unit (Bi-GRU), and a Conditional Random Field (CRF). Stage 2 applies a local span refinement module which revisits each candidate span to adjust both category assignments and boundary offsets. Our framework achieves a final score of 0.5668 which substantially outperforms the provided baseline. Ablation results further suggest that BiGRU-based contextual consolidation improves CRF-based decoding in text span extraction and classification.

1 Introduction

Discharge summaries in electronic health records provide a rich source of clinical information for decision-making. However, the heterogeneity of the text, semantic overlap of the decision categories, and uncertainty in identifying text boundaries corresponding to highly unbalanced labels make automatic extraction of text for decision support extremely challenging. The MedExACT (Elgaar et al., 2026) competition in BioNLP @ ACL 2026 formulates medical concept and text span extraction as a classification task over the MedDec dataset (Elgaar et al., 2024), the task requires systems to jointly detect decision spans and assign each span to one of nine DICTUM categories (Ofstad et al., 2016), while evaluating robustness of the approach across multiple demographic subgroups. This setting poses three compounding challenges as discharge summaries often exceed encoder context windows, DICTUM categories overlap semantically, and the stringent exact-match criteria render minor boundary offsets as complete span-level failures.

To address these challenges, we propose CASPAR, a two-stage approach to here. Stage 1 implements a pipeline with the deep learning backbone of a RoBERTa-base, a Bidirectional Gated Recurrent Unit (Bi-GRU), and a Conditional Random Field (CRF), to improve contextual representation learning for document-level span detection. To further improve the precision of span boundary detection, we introduce a lightweight span refinement module in Stage 2, which revisits each candidate span to correct both the category assignment and boundary offsets. Our system ranks 10th among the submissions from 37 competing teams (Final Score: 0.5668), substantially outperforming the organizer baseline. The source code for CASPAR is available at `https://github.com/schorm/caspar-medexact`.

2 Related Work

Recent NER research largely builds on token-level sequence labeling, where each token is assigned a BIO-, BILOU-, or related tag (Ramshaw and Marcus, 1995; Ratinov and Roth, 2009). In biomedical NER, this remains a central baseline formulation. Verma et al. (2023) organize mainstream biomedical NER approaches into simple token classifiers, CRF-based sequence labeling, and span

Proceedings of the BioNLP 2026 (Shared Tasks), pages 146–154

prediction. CRF-based decoders remain widely used to impose local label-transition constraints on top of contextual encoders (Huang et al., 2015; Jonker et al., 2024). In biomedical settings, token-level formulations also underpin strong practical systems such as AIONER (Luo et al., 2023), and recent cross-corpus tools such as HunFlair2 (Sänger et al., 2024). More recent works address dataset inconsistency and generalization issues within the sequence-labeling paradigm (Ruano et al., 2025).

Beyond token-level tagging, alternative formulations have been proposed to better model entity boundaries and span-level structure. Query-based and span-oriented methods include PIQN, which uses parallel instance queries for entity extraction (Shen et al., 2022), and BINDER, which models span–type matching through a contrastive bi-encoder formulation (Zhang et al., 2023). Two-stage and span-aware variants further emphasize boundary recovery including Locate and Label (Shen et al., 2021), document-level span fusion in ScdNER (Wei and Li, 2023), and boundary-aware generation in DiffusionNER (Shen et al., 2023). Tang et al. (2023) note that span-based methods can suffer from imbalanced span candidate spaces and difficulties in accurate boundary modeling.

Biomedical and clinical NER remains particularly challenging because texts are terminology-dense, annotation schemes often differ across datasets, and label spaces can be fine-grained and semantically overlapping. Recent work highlights these challenges from multiple angles, including annotation inconsistency across corpora (Ruano et al., 2025), large fine-grained label spaces (Yang et al., 2023), and the need to generalize beyond fixed entity inventories (Cocchieri et al., 2025). In medical decision extraction, these issues become even more pronounced. Elgaar et al. (2024) show that strong baselines still struggle under strict span-level evaluation; stronger token-level predictions do not necessarily translate into better exact-match span recovery, and adding CRF does not improve the baseline. These observations motivate our focus on contextual consolidation before CRF decoding and explicit second-stage span refinement.

3 Methodology

Figure 1 presents an overview of our proposed two-stage framework. Stage 1 performs document-level sequence labeling to detect candidate decision spans and assigns preliminary category labels. Stage 2 subsequently revisits each candidate span independently and refines its prediction using local context, correcting both category assignments and span boundaries. The two stages are trained separately, with Stage 2 taking the outputs of Stage 1 as its input.

The text span extraction and classification task is formulated as follows. Given a discharge summary D, the model predicts a set of labeled decision spans $S = \{(s_i, e_i, c_i)\}$, where s_i and e_i denote the start and end token indices of span i, and $c_i \in C$ denotes one of the nine following DICTUM categories: Contact related, Gathering information, Defining problem, Treatment goal, Drug related, Therapeutic procedure, Evaluating test result, Deferment, and Advice/precaution (Ofstad et al., 2016). The task therefore requires the model to jointly localize decision evidence in free text and assign each extracted span to the correct category. Following common practice in biomedical NER, we cast the problem as token-level sequence labeling with BIO tags over the label space

$$\mathcal{Y} = \{B\text{-}c, I\text{-}c, O \mid c \in C\}$$

where B-c marks the beginning of a span of category c, I-c marks its continuation, and O denotes tokens outside any decision span.

3.1 Stage 1: Document-Level Sequence Labeling

Discharge summaries are typically unstructured and often exceed the maximum input length supported by pretrained encoders. To preserve as much surrounding context as possible, we adopt a sliding-window strategy that partitions each document into overlapping chunks, which are encoded independently.

Given a discharge summary tokenized into a sequence of n tokens, $D = \{w_1, \ldots, w_n\}$, where w_i denotes the i-th token, we partition the document into M overlapping windows, denoted by $\{X^{(j)}\}_{j=1}^{M}$, where each chunk $X^{(j)}$ contains at most m tokens based on a fixed maximum window length shared across all windows. For each chunk, which is padded to length m when necessary, the encoder first produces contextualized token representations:

$$H^{(0,j)} = \text{Encoder}(X^{(j)}) \in R^{m \times d} \quad (1)$$

where d denotes the hidden dimensionality of the encoder.

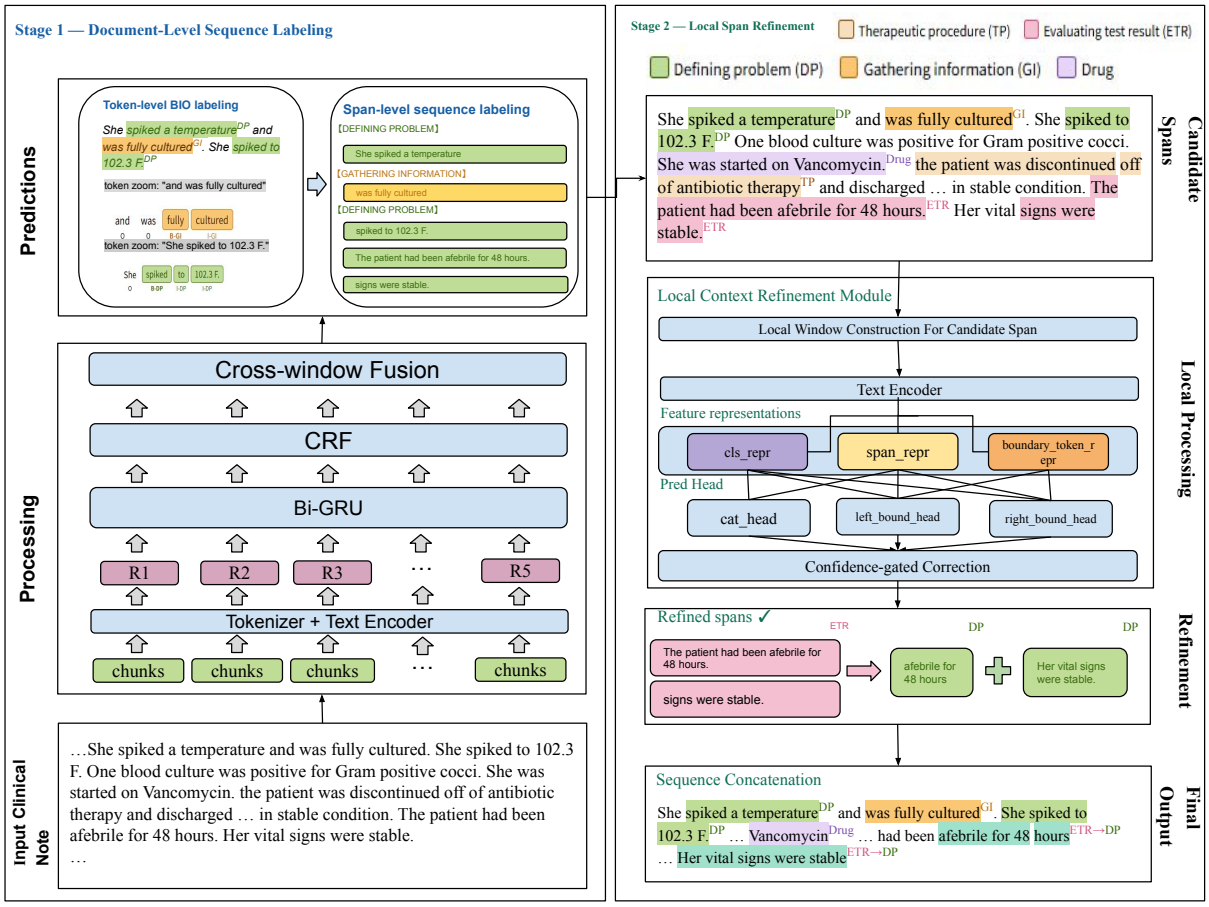

Figure 1: Architecture of CASPAR, our two-stage framework for medical decision span extraction and classification. Stage 1 performs document-level sequence labeling to produce candidate spans. Stage 2 applies local span refinement to correct category assignments and span boundary detection.

To further consolidate sequential dependencies before structured decoding, we pass these representations through a single-layer bidirectional GRU:

$$H^{(j)} = \mathrm{BiGRU}(H^{(0,j)}) \in R^{m \times d} \tag{2}$$

The resulting hidden states are then projected to token-level logits over the BIO label set $\mathcal{Y}$:

$$E^{(j)} = H^{(j)} W_e + b_e, \qquad E^{(j)} \in R^{m \times |\mathcal{Y}|} \tag{3}$$

where $W_e \in R^{d \times |\mathcal{Y}|}$ and $b_e \in R^{|\mathcal{Y}|}$ are learnable projection parameters.

Since each document token may appear in multiple overlapping windows, we reconstruct document-level predictions by fusing the emission logits across windows. Let w_t denote the t-th token in the original discharge summary D. Each window $X^{(j)}$ corresponds to a contiguous segment of the original document, covering token positions from a_j to b_j. Let $\mathcal{W}(t)$ denote the set of all windows that contain token w_t, i.e.

$$\mathcal{W}(t) = \{j \mid a_j \le t \le b_j\}. \tag{4}$$

For each such window $j \in \mathcal{W}(t)$, the local position of token w_t within window $X^{(j)}$ is defined as

$$p_j(t) = t - a_j + 1. \tag{5}$$

We then compute the fused emission logits for token w_t as

$$\tilde{E}_t = \frac{\sum_{j \in \mathcal{W}(t)} \alpha_{j,t} E^{(j)}_{p_j(t)}}{\sum_{j \in \mathcal{W}(t)} \alpha_{j,t}}. \tag{6}$$

Here, $\alpha_{j,t}$ is a triangular weight that attenuates predictions made near window boundaries. For a window of length m_j, with $\epsilon = 0.1$, center $c_j = (m_j+1)/2$, and radius $r_j = \max(c_j - 1,\ m_j - c_j)$, we define

$$\alpha_{j,t} = \max\left(\epsilon,\ 1 - \frac{|p_j(t) - c_j|}{r_j}\right). \tag{7}$$

Thus, predictions near chunk boundaries contribute less to the final document-level representation.

Finally, we apply a global CRF decoder over the fused logits to obtain the document-level BIO

tag sequence, where $\mathcal{Y}$ is the token-level BIO label space defined above, and $\mathcal{Y}^n$ denotes the set of all possible length-n BIO tag sequences

$$\hat{y} = \arg\max_{y \in \mathcal{Y}^n} \mathrm{Score}_{\mathrm{CRF}}(\tilde{E}, y). \tag{8}$$

Here, the CRF combines token-level emission scores with learned transition scores between adjacent labels, enabling globally consistent BIO decoding.

3.2 Stage 2: Local Span Refinement

Each Stage 1 candidate span $\hat{S}_i = (\hat{s}_i, \hat{e}_i, \hat{c}_i)$ is further processed by a lightweight local refinement module. Rather than re-encoding the full document, Stage 2 revisits each candidate in a local window. Given a document of length n tokens, we expand the predicted span by k context tokens on both left and right sides:

$$l_i = \max(1,\, \hat{s}_i - k), \qquad r_i = \min(n,\, \hat{e}_i + k) \tag{9}$$

The resulting token interval $[l_i, r_i]$ is mapped back to the original character offsets to extract the corresponding local text, which is then re-tokenized for the encoder model. The predicted span boundaries are aligned to the re-tokenized local sequence via the corresponding character offsets, yielding the local boundary indices p_i^L and p_i^R. This yields contextual token representations $H_i^{\mathrm{loc}} \in R^{L_i \times d}$, where L_i is the local sequence length after tokenization.

Followed the local processing module As shown in Fig. 1, from H_i^{loc}, we derive three feature representations as following a global context representation $\mathbf{h}_i^{\mathrm{cls}} = H_i^{\mathrm{loc}}[0]$ from `[CLS]` pooling; a span representation

$$\mathbf{h}_i^{\mathrm{span}} = \frac{1}{|M_i|} \sum_{t \in M_i} H_i^{\mathrm{loc}}[t], \tag{10}$$

obtained by mean-pooling over the predicted span tokens within the local window, where p_i^L and p_i^R denote the local token indices of the predicted left and right span boundaries, respectively, and

$$M_i = \{p_i^L, p_i^L + 1, \ldots, p_i^R\}$$

is the set of local token indices covered by the i-th predicted span; and a boundary representation

$$\mathbf{h}_i^{\mathrm{bdry}} = \left[H_i^{\mathrm{loc}}[p_i^L]\,;\, H_i^{\mathrm{loc}}[p_i^R]\right], \tag{11}$$

formed by concatenating the hidden states at the predicted left and right span boundaries.

These representations are consumed by a category head, two boundary heads, and two binary gate heads, where the gate heads predict whether a boundary correction should be applied on the left and right sides, respectively.

$$\mathbf{o}_i^{\mathrm{cat}} = W_{\mathrm{cat}}\left[\mathbf{h}_i^{\mathrm{cls}};\, \mathbf{h}_i^{\mathrm{span}}\right] + b_{\mathrm{cat}} \tag{12}$$

where $\mathbf{o}_i^{\mathrm{cat}} \in R^{|C|}$, $\mathbf{h}_i^{\mathrm{cls}} \in R^d$, $\mathbf{h}_i^{\mathrm{span}} \in R^d$, and $|C|$ is the number of decision categories.

The left and right boundary heads operate on $\mathbf{h}_i^{\mathrm{bdry}}$ and predict discrete offset classes over a label space of size $2\delta_{\max} + 1$:

$$\mathbf{o}_i^L = W_L \mathbf{h}_i^{\mathrm{bdry}} + b_L, \qquad \mathbf{o}_i^R = W_R \mathbf{h}_i^{\mathrm{bdry}} + b_R \tag{13}$$

where $\mathbf{h}_i^{\mathrm{bdry}} \in R^{2d}$, $W_L, W_R \in R^{(2\delta_{\max}+1)\times 2d}$, and $\delta_{\max}$ is the maximum allowable offset for boundary predictions.

To construct Stage 2 training instances, we run Stage 1 on the training split and retain each predicted span that has a token-level IoU > 0 with at least one gold span. For each retained candidate, the matched gold span (s_i, e_i, c_i) defines the boundary offsets

$$\Delta_i^L = s_i - \hat{s}_i, \qquad \Delta_i^R = e_i - \hat{e}_i \tag{14}$$

which are clipped to $[-\delta_{\max}, \delta_{\max}]$ to keep the prediction space tractable. The category head is optimized with label-distribution-aware margin loss (LDAM) (Cao et al., 2019) to mitigate class imbalance, while boundary and gate heads are trained with standard cross-entropy. The gate loss $\mathcal{L}_{\mathrm{gate}}$ is defined as the average of the left and right gate classification losses. The joint objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cat}} + \lambda_b \cdot \tfrac{1}{2}\left(\mathcal{L}_{\mathrm{left}} + \mathcal{L}_{\mathrm{right}}\right) + \lambda_g \cdot \mathcal{L}_{\mathrm{gate}}. \tag{15}$$

At inference time, category refinement is always applied, whereas boundary correction is applied only when both the boundary-head confidence and the gate probability exceed predefined thresholds τ_b and τ_g, respectively.

Denoting the accepted offsets as $\tilde{\Delta}_i^L$ and $\tilde{\Delta}_i^R$, the refined span is

$$s_i^* = \hat{s}_i + \tilde{\Delta}_i^L, \qquad e_i^* = \hat{e}_i + \tilde{\Delta}_i^R \tag{16}$$

Finally, refined spans are aggregated at the document level. Rather than score-based suppression, we merge spans of the same category that overlap or are directly adjacent, sorting by $(\text{category}, \text{start}, \text{end})$ and merging consecutive same-category spans whenever the next span starts no later than the end of the previous.

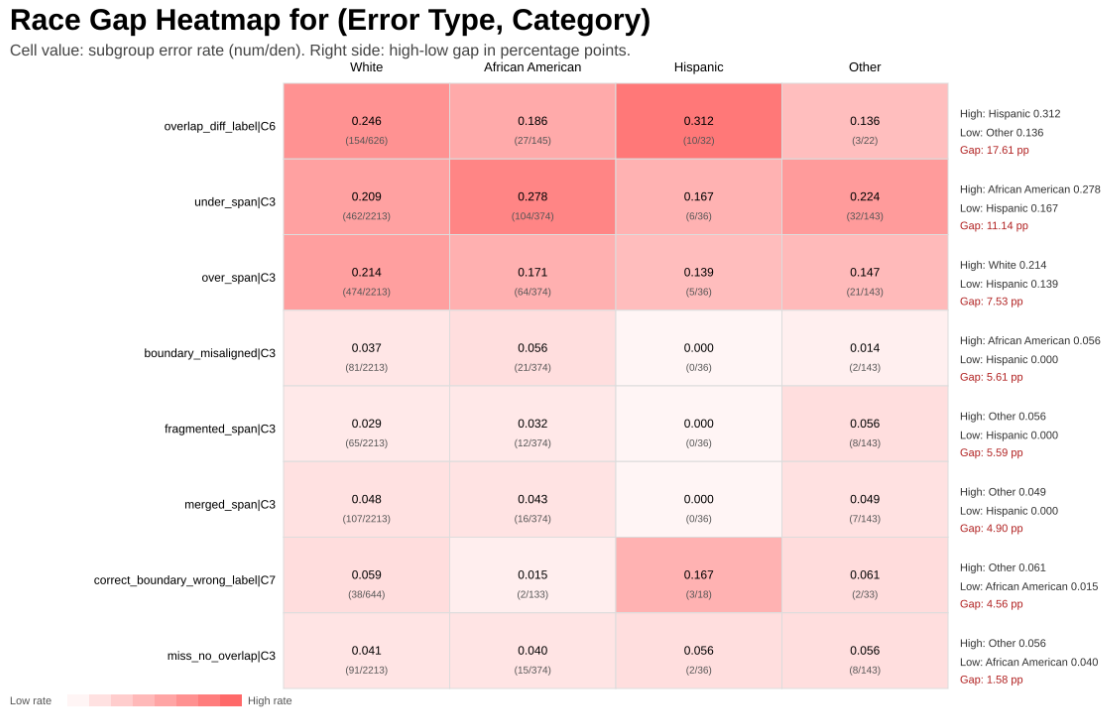

Figure 2: Error-type disparity heatmap across racial subgroups.

3.3 Implementation Details

We follow the official MedExACT shared-task data split (Elgaar et al., 2026) and use the provided training and validation partitions throughout the experiments. As gold labels for the test set remain proprietary to the task organizers, local evaluation is restricted to the validation split. Consequently, all reported test results are derived exclusively from the official leaderboard following system submission.

All experiments use RoBERTa-base (Liu et al., 2019) with m=512 maximum length window as the text encoder. Stage 1 is trained for 10 epochs with an encoder learning rate of 1×10^{-5}, a task-head learning rate of 5×10^{-5}, and a batch size of 8. Stage 2 uses an independently initialized RoBERTa-base encoder which is trained for 10 epochs using the same learning rates with a batch size of 32. The local window expands each candidate span by 32 tokens on both sides before re-tokenization, and the local encoder input is truncated to a maximum length of 128 tokens. Boundary offsets are discretized with $\delta_{\max} = 3$, yielding 7 offset classes per boundary head. On training-set matched candidates, ± 3 tokens cover 92.6%/93.2% of left/right boundary errors (86.8% jointly). The boundary loss weight λ_b is set to 1.0, the gate loss weight λ_g to 0.2, the boundary confidence threshold τ_b to 0.6, the gate probability threshold τ_g to 0.5, and the merge gap to 0.

3.4 Main Results

Table 1 reports the official MedExACT test results. CASPAR achieves a Final Score of 0.5668 on the shared-task leaderboard. For reference, we also report the mean and median scores across the 37 visible leaderboard submissions. Compared with the organizer baseline, CASPAR substantially improves Span F1 (0.5060 vs. 0.348).

In addition to the results from official leaderboard, Table 2 reports the repeated-run validation performance of CASPAR as mean $\pm$ standard deviation together with 95% confidence intervals which confirms the reliability of CASPAR and demonstrates that its gains are not artifacts of random initialization.

3.5 Ablation Study

Table 3 presents validation results across system variants. Adding BiGRU before CRF yields a substantial gain in Final Score (+8.0pp), suggesting that contextual consolidation makes CRF decoding more effective for boundary-sensitive prediction. This finding also helps explain why CRF alone provides only limited gains, consistent with the

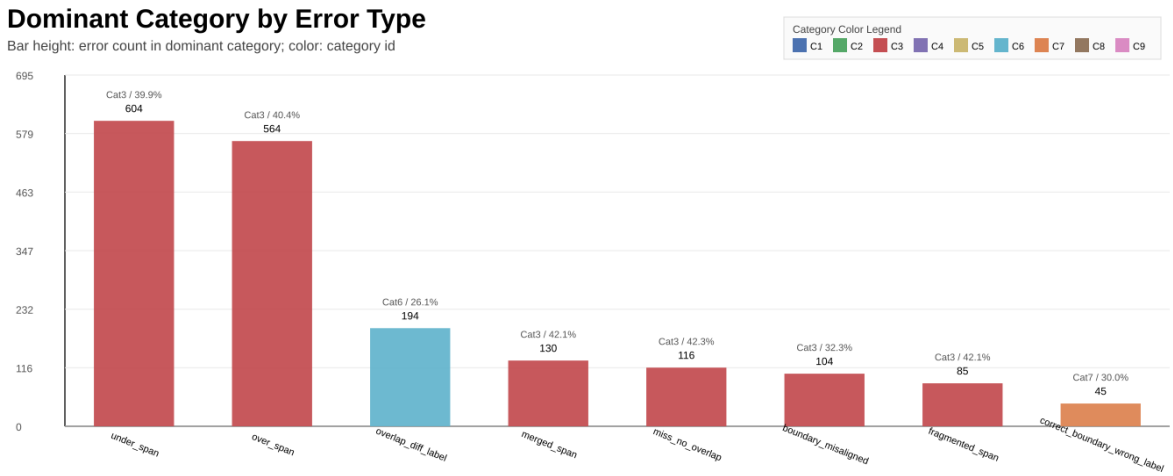

Figure 3: Dominant decision category by error type. Bar height shows the error count in the dominant category, and color indicates the category. Categories correspond to the nine DICTUM labels: Contact related (C1), Gathering information (C2), Defining problem (C3), Treatment goal (C4), Drug related (C5), Therapeutic procedure (C6), Evaluating test result (C7), Deferment (C8), and Advice/precaution (C9).

System	Final	WG	Span F1	Token F1
1st place	0.5965	0.5886	0.5419	0.6667
LB median	0.5331	0.5210	0.4605	0.6487
LB mean	0.5118	0.4983	0.4320	0.6193
Baseline†	–	–	0.348	–
CASPAR	0.5668	0.5602	0.5060	0.6409

Table 1: Official MedExACT test results. LB median = the Median of the results from MedEXact Leaderboard, LB mean = the mean of the results from MedEXact Leaderboard, Final = Final Score, WG = Worst-Group F1 which following the MedExACT evaluation metric.

Metric	Mean ± Std	95% CI
Final Score	0.5007 ± 0.0073	± 0.0064
WG Score (Hispanic)	0.4443 ± 0.0129	± 0.0113
BG Score (Female)	0.5759 ± 0.0010	± 0.0009
Span F1	0.4872 ± 0.0008	± 0.0007
Token F1	0.6271 ± 0.0040	± 0.0035

Table 2: Run-to-run stability of CASPAR measured over five validation runs using different random seeds (42, 13, 21, 84, and 100), and where WG = Worst-Group and BG = Base-Group.

observations of Elgaar et al. (2024). Stage 2 refinement further improves prediction performance, with a more modest gain in the Final Score.

4 Discussion

Figure 2 shows that racial subgroup disparities are associated with different error patterns. Hispanic patients exhibit the highest label confusion rate (31.69%), mainly between semantically overlapping categories such as *Therapeutic Procedure* (cat 6) and *Defining Problem* (cat 3). For example, *"NG lavage with coffee ground and brown clot"* is repeatedly predicted as cat 3 despite being annotated as cat 6. African American patients show the highest under-span rate in cat 3 (27.8%), where clinically meaningful qualifiers are truncated, as in *"AFib on coumadin"* being reduced to *"AFib"*. By contrast, the *Other* group has the highest miss rate (6.35%), indicating a stronger recall problem. Overall, these results suggest that subgroup disparities in this task arise from different combinations of category ambiguity, boundary truncation, and recall loss.

System	Final	Span F1	Token F1
RoBERTa (only)	0.4359	0.3939	0.5601
RoBERTa + CRF	0.4378	0.4065	0.5757
+ BiGRU (Stage 1)	0.4649	0.4396	0.5961
+ Stage 2 (full)	0.4998	0.4873	0.6249

Table 3: Ablation results based on the validation set of MedExACT competition benchmark.

Additionally, From Table. 1, the gap between token F1 score (0.64) and span F1 score (0.51) highlights a structural mismatch between token-level BIO modeling and exact-match span evaluation. As shown in Figure 3, under-span and over-span errors dominate (24.21% and 19.88%, respectively), while label confusion (`overlap_diff_label`) is concentrated in cat 6 (26.1%). As a result, even small boundary deviations such as retaining a trail-

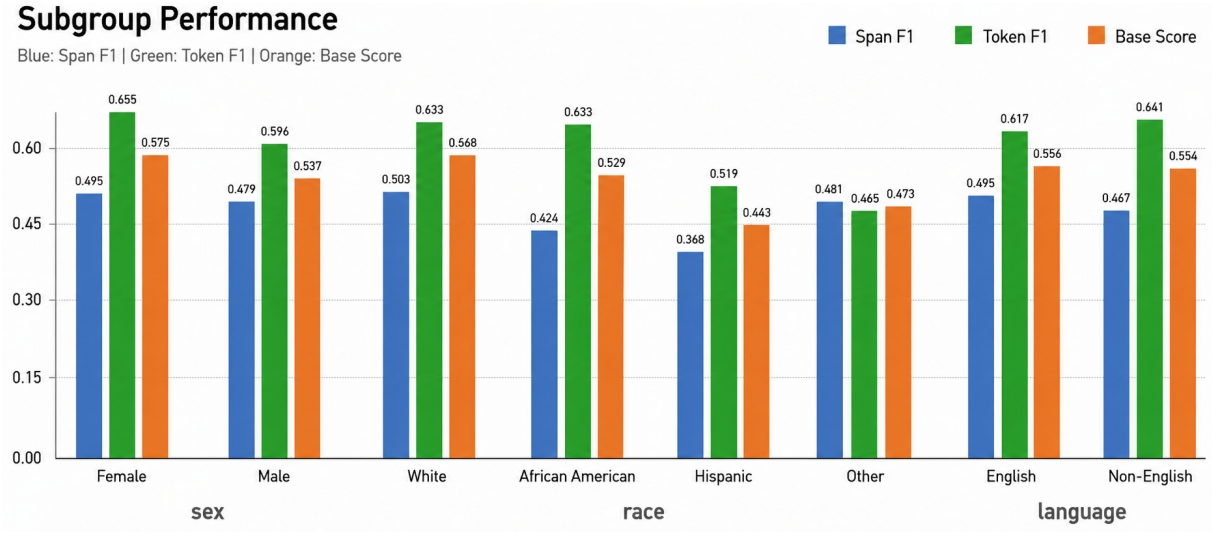

Figure 4: Subgroup-level performance across demographic groups. Token F1 consistently exceeds Span F1 for all subgroups, with the largest gap observed in Hispanic patients (Token F1: 0.519 vs. Span F1: 0.368), reflecting the challenge of exact span boundary recovery under demographic distribution shift.

ing punctuation mark or dropping a leading determiner lead to complete span-level errors. This is partly due to the BIO formulation itself, predicting $2|\mathcal{C}| + 1$ labels over a multi-class tag space increases task complexity and offers only indirect supervision for recovering span boundaries as complete units. The reasons for this observation is that the model frequently suffers from boundary expansion and category drifting. For instance, in detecting cat 6 co-occurring are often merged medications (e.g., swallowing "combivent neb" into a preceding drug phrase) or procedures are not properly isolated from their narrative context (e.g., expanding "CTA" into the entire event description). These cases demonstrate that while the model effectively locates the general vicinity of *Therapeutic Procedure* (cat 6) entities, it fails to filter out redundant contextual noise. Consequently, the model frequently conflates core denoting medical tokens procedure with adjacent narrative descriptions or co-occurring phrases, leading to imprecise boundaries and degraded span-level performance.

The CASPAR pipeline relies on BIO-based token-level prediction for span detection, which introduces an inherent mismatch with strict span-level evaluation. As shown in Figure 4, strong performance at the token level does not necessarily translate to precision in span recovery based on exact-match (EM) criteria. The *Non-English* subgroup achieves the highest Token F1 (0.641) but only a moderate Span F1 (0.467), while the *Other* subgroup attains comparable span-level performance despite much weaker token-level predictions. It suggests that token-level supervision alone is insufficient for fully capturing precise span boundaries. This discrepancy arises primarily because most failed recoveries are not due to missed detections, but rather flawed span boundaries. As illustrated by the *Defining problem (C3)* category in Figure 3, which exhibits the highest rate of under-span errors, the model tends to prioritize high-confidence keywords that define the problem while neglecting the complete boundary of the span.

A complementary limitation arises from the design of Stage 2. Because training instances are constructed from Stage 1 predictions that have token-level IoU greater than 0 with a gold span, the refinement module is trained exclusively on partially correct candidates. As a result, Stage 2 cannot explicitly suppress false-positive spans produced by Stage 1, nor can it recover spans that Stage 1 missed entirely. This structural constraint provides a principled explanation for why Stage 2 yields meaningful but modest gains in the ablation study (Table 3). It improves boundary precision and category assignment within the set of retrieved candidates, but leaves recall errors inherited from Stage 1 unaddressed. A natural direction for future work is to incorporate false-positive rejection and span proposal generation directly into the refinement stage, so that Stage 2 can correct a broader range of Stage 1 errors.

5 Conclusion

We present a two-stage pipeline for medical decision span extraction and classification. Stage 1 combines a sliding-window text encoder with BiGRU and CRF to produce document-level candidate spans, while Stage 2 applies a lightweight local refinement module to correct category assignments and span boundaries per candidate. Our system achieved a final score of 0.5668. Results from the ablation study confirm that BiGRU consolidation is critical for effective CRF decoding. The error analysis shows that demographic subgroup disparities are driven by different combinations of category confusion and boundary truncation. Moreover, the persistent gap between the token-level and span-level performance highlights an inherent limitation of BIO-based formulations under exact-match evaluation. A natural direction for future work is to incorporate span-aware objectives directly into training, so that the model can optimize token classification while simultaneously learning to recover exact span boundaries more reliably. Beyond the Stage 1 formulation, extending Stage 2 to handle a broader range of candidate types, including false-positive rejection and recovery of spans missed by Stage 1 would address the structural recall ceiling identified in our analysis and potentially yield more substantial gains from the refinement stage.

References

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Alessandro Cocchieri, Viet Dac Lai, Sukannya Purkayastha Boro, and et al. 2025. Openbioner: Lightweight open-domain biomedical named entity recognition through entity type description. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 847–860. 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.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence labeling. *arXiv preprint arXiv:1508.01991*.

Richard A. A. Jonker, Tiago Almeida, Rui Antunes, João R. Almeida, and Sérgio Matos. 2024. Multi-head crf classifier for biomedical multi-class named entity recognition on spanish clinical notes. *Database*, 2024:baae068.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu. 2023. Aioner: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics*, 39(5):btad310.

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.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In

Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), pages 147–155. Association for Computational Linguistics.

João Ruano, Gonçalo Correia, Leonor Barreiros, and Afonso Mendes. 2025. Effective multi-task learning for biomedical named entity recognition. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 225–239, Viena, Austria. Association for Computational Linguistics.

Mario Sänger, Alan Akbik, Amir Pouran Ben Veyseh, and et al. 2024. Hunflair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools. *Bioinformatics*, 40(10):btae564.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and label: A two-stage identifier for nested named entity recognition. 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 2782–2794. Association for Computational Linguistics.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Diffusioner: Boundary diffusion for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7186–7201. Association for Computational Linguistics.

Yongliang Shen, Xiaobin Wang, Zeqi Tan, Guangwei Xu, Pengjun Xie, Fei Huang, Weiming Lu, and Yueting Zhuang. 2022. Parallel instance query network for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 947–961, Dublin, Ireland. Association for Computational Linguistics.

Minghao Tang, Yongquan He, Yongxiu Xu, Hongbo Xu, Wenyuan Zhang, and Yang Lin. 2023. A boundary offset prediction network for named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14834–14846, Singapore. Association for Computational Linguistics.

Harsh Verma, Sabine Bergler, and Narjesossadat Tahaei. 2023. Comparing and combining some popular NER approaches on biomedical tasks. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 273–279, Toronto, Canada. Association for Computational Linguistics.

Ying Wei and Qi Li. 2023. Scdner: Span-based consistency-aware document-level named entity recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15677–15685. Association for Computational Linguistics.

Jiuding Yang, Jinwen Luo, Weidong Guo, Di Niu, and Yu Xu. 2023. Exploiting hierarchically structured categories in fine-grained chinese named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3407–3421, Toronto, Canada. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2023. Optimizing bi-encoder for named entity recognition via contrastive learning. In *The Eleventh International Conference on Learning Representations*.