ELiRF-UPV@MedExACT 2026: Dynamic Section Conditioning for Medical Decision Span Detection in Discharge Summaries

Vicent Ahuir[†], **Lluís-F. Hurtado**[†,§], **María-José Castro-Bleda**[†,§]
[†]VRAIN: Valencian Research Institute for Artificial Intelligence
Universitat Politècnica de València, Spain
[§]ValgrAI: Valencian Graduate School and Research Network of Artificial Intelligence
[†]vahuir@upv.es, lhurtado@upv.es, mcastro@dsic.upv.es

Abstract

Extracting medical decisions from discharge summaries is essential for downstream clinical analytics, yet the task remains challenging due to the heterogeneous structure of electronic health records. For the MedExACT track at ACL 2026, we proposed a system that achieved the 4th position. Our approach first applies dynamic section conditioning to capture the contextual dependencies inherent in each document. A transformer backbone is then augmented with category- and section-aware layer mixing, enabling us to fuse global document structure with fine-grained semantic cues. To further improve robustness, we employ an ensemble of instruction-tuned large language models for automatic section extraction, while a fairness-oriented model selection criterion ensures that performance does not degrade on minority demographic subgroups. The resulting system attains a final score of 0.5806 on the held-out test set and demonstrates significant gains over the baseline across all evaluated subpopulations.

1 Introduction

The extraction of medical decisions from clinical text is a critical task that holds significant potential for enhancing patient care, advancing medical research, and developing predictive models (Oliveira et al., 2025; Klug et al., 2024). However, this process is fraught with challenges due to the inherent complexity of clinical text, which is characterized by specialized jargon, abbreviations, and diverse formats such as notes and reports (Mortadi et al., 2025; Ando et al., 2022). Discharge summaries, in particular, are rich repositories of medical decisions, yet their extraction remains non-trivial due to the inherent complexity and contextuality of clinical language (Navarro et al., 2023; Hossain et al., 2023).

The MedExACT task at ACL 2026 (Elgaar et al., 2026) builds on this challenge by requiring systems to identify and classify medical decisions within ICU discharge summaries. The task is grounded in the MedDec dataset (Elgaar et al., 2024), which comprises over 56 000 expert-annotated spans derived from de-identified MIMIC-III clinical notes (Johnson et al., 2016) across nine decision categories. This dataset reflects real-world clinical practice, with significant variations in demographic and linguistic distributions, making it a robust benchmark for evaluating both accuracy and equity in medical decision extraction systems.

This task faces significant challenges: contextual complexity, demographic variability, and overlapping semantics. Contextual complexity demands models that capture fine-grained semantic dependencies across document sections. Demographic variability requires consistent performance across diverse groups to ensure fairness. Overlapping semantics necessitate robust methods to accurately disambiguate decisions.

Our paper addresses these issues through an innovative approach involving dynamic section conditioning and the integration of semantic segmentation with context-aware modeling. This method not only enhances the accuracy of medical decision extraction but also ensures fairness across diverse populations.

The remainder of the paper is structured as follows: Section 2 provides an overview of the task, Section 3 details our methodology, Section 4 presents our experimental results, and Section A describes our validation findings. Finally, Section 6 discusses the implications of our findings and future directions.

2 The Shared Task

The MedExACT shared task was guided by the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM). Participating systems were required to detect contiguous

Proceedings of the BioNLP 2026 (Shared Tasks), pages 71–76

Table 1: Gold span counts and percentage distribution per demographic subgroup across dataset splits in the MedDec dataset (Elgaar et al., 2024). Test counts (n) are derived from the shared-task evaluation output.

		Train		Eval		Test	
Dimension	**Group**	n	%	n	%	n	%
Sex	Female	17342	42.1	3262	49.7	2298	37.8
	Male	23831	57.9	3307	50.3	3780	62.2
Race	White	28342	68.8	5159	78.5	4673	76.9
	African American	4078	9.9	902	13.7	313	5.1
	Hispanic	2670	6.5	175	2.7	89	1.5
	Asian	1154	2.8	0	0.0	148	2.4
	Other	4929	12.0	333	5.1	855	14.1
Language	English	25762	62.6	4744	72.2	4134	68.0
	Non-English	15411	37.4	1825	27.8	1944	32.0

text spans and classify them into nine categories: *Contact related*, *Gathering information*, *Defining problem*, *Treatment goal*, *Drug*, *Therapeutic procedure*, *Evaluating test result*, *Deferment*, and *Advice/precaution*.

Table 1 shows how gold span counts vary by sex, race, and language across the three splits in the MedDec dataset. In the training set, males dominate (57.9%), but the evaluation split is almost balanced (50.3%), and the test set again skews male (62.2%). Racial composition shifts as well: White subjects rise from 68.8% in training to 76% in both evaluation and test, while African-American spans peak at 13.7% during evaluation and fall to 5.1% in the test set. Language distribution mirrors this trend, with English spans peaking at 72.2% in evaluation and non-English spans being highest in training (37.4%).

In the shared task, the systems were evaluated with a composite metric that balances accuracy and fairness. Accuracy is captured by the Base Score S_{base}, the arithmetic mean of Span-F_1 (exact span-and-label match) and Token-F_1. Fairness is measured via the Worst-group Score $S_{\text{worst}} = \min_{g \in G} S_{\text{base},g}$, where G denotes all demographic or linguistic subgroups. The Final Score, encouraging uniform performance across groups, is

$$S_{\text{final}} = \frac{S_{\text{base}} + \min_{g \in G}(S_{\text{base},g})}{2}.$$

The baseline was a RoBERTa-based BIO sequence tagger. Because clinical notes average 1600 words, the model processed them in 512-token chunks to respect input limits (Elgaar et al., 2024).

3 Developed System

The developed system comprises a modular pipeline that integrates semantic segmentation of clinical reports with a context-aware extraction phase. The architecture employs a pretrained transformer encoder to generate a sequence of hidden states across all layers; subsequently, a category- and section-aware layer-mixing module aggregates these states to produce the final embeddings for the classification heads. By utilizing an attention mechanism, the layer-mixing module computes a weighted sum of the encoder's hidden states, conditioned on the section's canonical identity and the targeted entity type. This configuration is designed to facilitate the integration of global document structure with local semantic features, providing a framework to manage the contextual dependencies inherent in segmented clinical text.

3.1 Section Extraction

Clinical reports often lack formal partitioning, leading to segmentation strategies that ignore natural semantic boundaries. However, clinical discourse is fundamentally structured around distinct headers –such as *History*, *Findings*, and *Impression*– which provide critical inductive biases for entity contextualization. For instance, an entity labeled as a condition in the *History* section implies a past state, whereas the same entity in *Findings* denotes a current observation.

To overcome the brittleness of rule-based parsing, this system delegates section identification to an ensemble of instruction-tuned Large Language Models (LLMs), including *Llama 3.2* (Grattafiori et al., 2024), *Gemma 3* (Team et al., 2025), *Qwen3* (Yang et al., 2025), and *Mistral* (Jiang et al., 2023). These models are leveraged to resolve the stylistic variations and semantic synonymy common in free-text reports. We implement a sequential inference strategy to maintain computational efficiency,

terminating the ensemble query once a valid segmentation is achieved. This process transforms unstructured text into structured records characterized by canonical names and precise character offsets.

To ensure the reliability of these extractions, we utilize a multi-dimensional scoring framework. This metric evaluates the completeness of text coverage, penalizes overlapping segments, and validates that section lengths remain within clinically plausible bounds. By applying a statistical penalty to high variance in segment distribution, the system prioritizes structural consistency, ensuring that internally balanced segmentations are selected over irregular partitions.

3.2 Architecture

To handle the presence of tokens belonging to multiple semantic categories, the architecture employs 9 independent BIO classification heads operating in parallel, one per category. This parallelized structure enables the representation of overlapping spans without specialized decoding, as each head independently predicts the presence of a specific entity type. The system's training objective is formulated as the mean of the 9 per-head Conditional Random Field (CRF) negative log-likelihood losses.

To provide structured conditioning, section names are clustered with the Hierarchical Density-Based Spatial Clustering of Applications with Noise, HBSCAN, (Campello et al., 2013) into semantically coherent groups using the `all-MiniLM-L6-v2` model of SentenceTransformer (Reimers and Gurevych, 2019). During training, headers are grouped into clusters of at least 3; headers in the validation and test sets are subsequently mapped to their nearest training cluster using cosine similarity. This establishes a fixed vocabulary of canonical section identities, allowing the model to incorporate document-level context despite variations in header nomenclature.

Documents are processed at the section level, with each section treated as an independent training instance. Sections exceeding the transformer's token capacity are partitioned into overlapping windows of 768 characters with a 50-character stride. Annotations are aligned to sub-tokens via offset mapping, with overlapping boundaries resolved greedily to maximize character coverage. Those parts of the text that are not covered by any section are assigned to a dedicated "undefined" section.

The architecture utilizes a pretrained transformer backbone to extract a sequence of hidden states H. A subsequent layer-mixing module computes a weighted combination of all $L+1$ hidden states. Using an attention mechanism, the model selects layers based on both the category type and the specific section identity. For a given head, we define a 2-token query sequence $Q = [e_{category}, e_{section}]$, where $e_{section}$ is derived from the section cluster embedding and $e_{category}$ from one of the nine category embeddings. The attention weights α for each layer $l \in \{0, \ldots, L\}$ are computed by first calculating the alignment scores:

$$a_{b,s,l} = \frac{1}{2}((W_k H_{b,s,l}) \cdot (W_q e_{category}) + (W_k H_{b,s,l}) \cdot (W_q e_{section}))$$

$$\alpha_{b,s,l} = \frac{\exp(a_{b,s,l})}{\sum_{j=0}^{L} \exp(a_{b,s,j})}$$

The resulting conditioned representation $h'_{b,s}$ is the weighted sum of the previously extracted hidden states:

$$h'_{b,s} = \sum_{l=0}^{L} \alpha_{b,s,l} \cdot H_{b,s,l}$$

Final sequence labeling is performed by the 9 labeling heads, each containing a specialized Bidirectional LSTM (BiLSTM) and a CRF layer. This architecture ensures that each head models category-specific sequential patterns.

4 Submitted Runs

We submitted three runs, differing only in encoder, initialization, and model-selection strategy; all were fine-tuned on the shared-task data (LR = 5×10^{-5}, 10% linear warm-up over 5 epochs unless noted).

- **Run 1**: Bio-ClinicalBERT (Alsentzer et al., 2019) encoder; checkpoint chosen to maximize the overall competition score on validation.
- **Run 2**: RoBERTa-large (Zhuang et al., 2021) encoder; same fine-tuning and selection criterion as Run 1, allowing a direct size and domain-vs-general comparison.
- **Run 3**: The fine-tuning process of Run 2, but changing the checkpoint selection to the maximization of the harmonic mean across demographic subgroups to prioritize fairness over aggregate performance.

Validation results are presented in Appendix A.

5 Shared Task Results

Table 2 reports the overall performance of our three runs on the MedExACT @ ACL2026 test set, while Table3 gives a fine-grained analysis by demographic subgroup. The results show a clear improvement in both accuracy and stability over validation, indicating strong generalization to unseen clinical data.

Table 2: Results for the three submitted runs in the MedExACT @ ACL 2026 Test set.

Metric	Run 1	Run 2	Run 3
Span F1	0.4952	0.5174	0.5237
Token F1	0.6181	0.6610	0.6541
Worst-Group Score	0.5397	0.5650	0.5723
Final Score	0.5482	0.5771	0.5806

Table 2 shows that Run 3 is our best submission, with a *Final Score* of 0.5806 and a *Worst-Group Score* of 0.5723. In contrast to the validation phase (Appendix A), where *RoBERTa-large* lagged behind domain-specific *Bio-ClinicalBERT*, the test results favor the larger general-domain model. Thus, while Run 1 (*Bio-ClinicalBERT*) establishes a strong baseline, Run 3's scaling and fairness-oriented selection deliver superior overall performance and demographic equity on the test set.

Table 3: Subgroup base scores for the three submitted runs in the MedExACT @ ACL 2026 Test set.

Group	Subgroup	Run 1	Run 2	Run 3
Sex	Female	0.5816	0.6118	0.6149
	Male	0.5413	0.5757	0.5723
Race	White	0.5568	0.5876	0.5778
	African American	0.5419	0.5843	0.5966
	Hispanic	0.5454	0.5650	0.6134
	Asian	0.5460	0.5764	0.5860
	Other	0.5601	0.5988	0.6376
Language	English	0.5670	0.5980	0.5912
	Non-English	0.5397	0.5739	0.5837
Harmonic Mean		0.5530	0.5854	0.5964

Table 3 shows that the bottleneck for the *Hispanic* cohort seen in validation is largely eliminated on the test set: Run 3 scores 0.6134, surpassing many other groups. This confirms that optimizing the *Harmonic Mean* during checkpoint selection balances sensitivity across demographics, and that Run 3 also attains the highest overall *Harmonic Mean* (0.5964).

Table 4: Top 5 participants on the MedExACT @ ACL 2026 test leaderboard. Subscripts indicate the rank position for each metric among the five participants. The baseline (BL) is shown for reference. The columns refer to: Final Score (F-S), Baseline Score (B-S), Worst-Group Score (WG-S), Span F1 (S-F1), and Token F1 (T-F1).

#	F-S	B-S	WG-S	S-F1	T-F1
1	0.5965_1	0.6043_1	0.5886_1	0.5419_1	0.6667_3
2	0.5942_2	0.6003_2	0.5881_2	0.5257_2	0.6750_2
3	0.5809_3	0.5924_3	0.5695_4	0.5181_4	0.6666_4
*4	0.5806_4	0.5889_4	0.5723_3	0.5237_3	0.6541_5
5	0.5724_5	0.5848_5	0.5601_5	0.4900_5	0.6796_1
BL	0.5111	0.5301	0.4922	0.4363	0.6238

Table 4 places our best submission (Run 3) 4th overall, surpassing the official baseline by $+0.0695$ in *Final Score* and $+0.0801$ in *Worst-Group Score*. Although we rank 5th on Token-F1, our 3rd-place Worst-Group Score underscores the model's effectiveness at achieving robust, equitable clinical entity extraction.

6 Conclusions

We demonstrated that dynamic section conditioning—semantic segmentation of clinical documents followed by a context-aware encoder—substantially improves medical decision extraction while maintaining demographic equity. Across three experimental runs, the RoBERTa-large model with a fairness-focused selection strategy (Run 3) achieved the highest Final Score (0.5806) and Worst-Group Score (0.5723), outperforming both the baseline and other configurations. The improvement from validation to test data confirms strong generalization to unseen clinical records.

Maximizing the harmonic mean across demographic subgroups in Run 3 not only boosted overall performance but also ensured equitable outcomes, illustrating that fairness-aware model selection can enhance both accuracy and inclusivity.

Future work will focus on reducing computational complexity (e.g., via distillation or quantization) and extending robustness evaluations to diverse clinical settings, thereby increasing the real-world applicability of our approach. In addition, we plan to conduct a thorough ablation study to better understand the contribution of key components.

This will include analyzing the impact of varying the number of retrieved similar sections on final performance, as well as systematically removing or varying components such as section conditioning, ensemble models, and layer mixing.

Limitations

While our approach shows strong performance overall, several limitations should be noted.

First, using multiple large language models (LLMs) for section extraction creates significant computational overhead—each model must run separately, increasing latency and GPU memory demands, which hampers scalability in resource-constrained clinical settings.

Second, the method assumes reliable section identification. In practice, documents vary widely in formatting and terminology; misidentified sections can propagate errors to downstream tasks such as entity or relation extraction.

Third, although promising on MedDec, performance on other clinical domains (radiology, pathology, and outpatient notes) has not yet been evaluated. Different specialties use distinct vocabularies and structures that may challenge both the heuristics and the LLMs' adaptability without fine-tuning.

Finally, reliance on large pre-trained models raises privacy and regulatory concerns. Even with local inference, model parameters can contain sensitive patterns from training data, requiring careful audit and mitigation before deployment in regulated environments.

Acknowledgments

This research is supported by Grant PID2024-155948OB-C55 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Kenichiro Ando, T. Okumura, Mamoru Komachi, H. Horiguchi, and Yuji Matsumoto. 2022. Is artificial intelligence capable of generating hospital discharge summaries from inpatient records? *PLOS Digital Health*, 1.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Elias Hossain, R. Rana, N. Higgins, J. Soar, P. Barua, Anthony R. Pisani, and K. Turner. 2023. Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review. *Computers in biology and medicine*, 155:106649.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Alistair Johnson, Tom Pollard, and Roger Mark. 2016. MIMIC-III Clinical Database. *PhysioNet*. Version 1.4.

Katrin Klug, Katharina Beckh, Dario Antweiler, Nilesh Chakraborty, Giulia Baldini, Katharina Laue, R. Hosch, F. Nensa, Martin Schuler, and Sven Giesselbach. 2024. From admission to discharge: a systematic review of clinical natural language processing along the patient journey. *BMC Medical Informatics and Decision Making*, 24.

Ahmad Mortadi, Waleed Nazih, Mohamed I. Eldesouki, and Yasser Hifny. 2025. Intelligent de-identification of medical discharge summaries using hybrid nlp

techniques. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24:1 – 17.

David Fraile Navarro, K. Ijaz, Dana Rezazadegan, H. Rahimi-Ardabili, M. Dras, E. Coiera, and S. Berkovsky. 2023. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International journal of medical informatics*, 177:105122.

Juliana Damasio Oliveira, H. D. P. Santos, Ana Helena D. P. S. Ulbrich, Julia Colleoni Couto, Marcelo Arocha, Joaquim Santos, Manuela Martins Costa, Daniela Faccio, F. Tabalipa, and Rodrigo F. Nogueira. 2025. Development and evaluation of a clinical note summarization system using large language models. *Communications Medicine*, 5.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Validation Results

The validation results for the three submitted runs are presented in Table 5, showing the overall performance of the models. The *RoBERTa-large* configurations (Run 2 and Run 3) yielded higher Span F1 and Token F1 scores than the domain-specific *Bio-ClinicalBERT* (Run 1), yet Run 1 reached the highest Final Score of 0.4732. The *Worst-Group Score* was highest in Run 1 (0.4089), suggesting that the domain-specific representations in this run maintain a higher performance floor on the validation set compared to the larger general-domain models.

Table 5: Validation set results for the three submitted runs.

Metric	**Run 1**	**Run 2**	**Run 3**
Span F1	0.4802	0.5013	0.5013
Token F1	0.5947	0.6153	0.6294
Worst-Group Score	0.4089	0.3772	0.3538
Final Score	0.4732	0.4677	0.4596

Table 6 provides the breakdown of scores across sex, race, and language subgroups. It shows that selecting for the Harmonic Mean in *Run 3* results in the highest consistency across most categories, notably improving the *Other* race subgroup to 0.5129. However, this optimization for parity across the majority of groups coincided with the lowest observed score for the *Hispanic* subgroup (0.3538), which remained the performance bottleneck across all three runs. While Run 3 achieved the highest overall Harmonic Mean (0.5221), the validation data suggests a trade-off: increasing broad demographic consistency could lower the score of the lowest-performing subgroup.

Table 6: Validation set subgroup base scores for the three submitted runs.

Group	**Subgroup**	**Run 1**	**Run 2**	**Run 3**
Sex	Female	0.5617	0.5819	0.5861
	Male	0.5138	0.5353	0.5452
Race	White	0.5469	0.5673	0.5726
	African American	0.5429	0.5712	0.5739
	Hispanic	0.4089	0.3772	0.3538
	Other	0.4109	0.4419	0.5129
Language	English	0.5250	0.5512	0.5577
	Non-English	0.5639	0.5736	0.5803
Harmonic Mean		0.5014	0.5138	0.5221