Overview of the Medical Decision Extraction, Analysis, and Classification Task (MedExACT) of BioNLP 2026

Mohamed Elgaar,[1] Jiali Cheng,[1] Nidhi Vakil,[1] Mehrnaz Sadrolashrafi,[2]
Mitra Mohtarami,[3] Adrian Wong,[2] Hadi Amiri,[1,4] Leo A. Celi[4]
[1]UMass Lowell, [2]BIDMC, [3]Anselm College, [4]MIT
{melgaar,jcheng2,nvakil,hadi}@cs.uml.edu, {msadrola,awong7}@bidmc.harvard.edu
mmohtarami@anselm.edu, lceli@mit.edu

Abstract

This paper presents an overview of the Medical Decision Extraction, Analysis, and Classification task (MedExACT) of BioNLP 2026. The focus of this task is the extraction and labeling of medical decisions in ICU discharge summaries. The task is built on MedDec, a MIMIC-III-based dataset of 451 expert-annotated summaries, and asks systems to extract and classify spans of text that contain medical decisions according to the decision categories defined in the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM). The official ranking combines span F1 and token F1 with a worst-group robustness metric computed over sex, race, and English-proficiency subgroups. MedExACT attracted broad international interest, with 130 official submissions from 36 teams comprising about 60–100 participants, and has improved information extraction performance by nearly 15% over the previous state of the art. The submitted systems predominantly use long-context encoder models, ensemble decoding, boundary-refinement modules, and robustness-aware training or model selection, with the best submitted run reaching a final fairness-based F1 of 0.596.

1 Introduction

Clinical notes often document key medical decisions concerning diagnoses, medications, procedures, follow-up plans, and other decisions that shape downstream care. These decisions are central to clinical reasoning, but they are embedded in long free-text narratives that make them difficult to analyze at scale.

Prior studies have investigated joint span detection and classification for clinical outcomes (Abaho et al., 2021), weakly supervised extraction of medical evidence from clinical texts (Liu et al., 2021), and structured event extraction using sequence-to-sequence frameworks (Ma et al., 2023).

MedDec (Elgaar et al., 2024) introduced the first expert-annotated benchmark focused on extracting medical decisions from discharge summaries, grounding the task in the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM) (Ofstad et al., 2016) and in the MIMIC-III critical care corpus (Johnson et al., 2016). The same data has also supported interactive downstream tooling for extraction, visualization, and annotation of medical decisions (Elgaar et al., 2025).

MedExACT 2026 extends the above efforts into a BioNLP shared task. Participants are asked to identify contiguous decision spans in ICU discharge summaries and assign each span one of the shared-task labels. Unlike other valuable information extraction resources such as CLIP (Mullenbach et al., 2021), which focus on physician action items, MedExACT targets a decision taxonomy that includes diagnostic assessments, therapeutic plans, and evaluation statements. The task averages standard performance (span- and token-level F1s) with a worst-group robustness metric across nine subgroups defined by sex, race, and English proficiency. This explicitly incentivizes fairness-aware system design instead of treating fairness as a post-hoc diagnostic. The task drew 130 submissions from 36 teams. The winning system scored 0.596. Top-performing approaches relied on encoder-based models with ensembles, boundary refinement, and category-specific decoding. LLMs performed well as augmentation or reranking components but poorly as standalone span extractors, consistent with prior MedDec findings.

This overview paper documents the task formulation and official evaluation approach, introduces the public baseline and toolkit distributed with the task, summarizes the systems submitted by participating teams and the modeling patterns that emerged from those submissions, and analyze the systems on exact span extraction, long-document processing, and subgroup-aware evaluation.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 760–769

Decision Category	Description	Examples
Contact related	Decision regarding admittance or discharge from hospital, scheduling of control and referral to other parts of the healthcare system	Admit, discharge, follow-up, referral
Gathering information	Decision to obtain information from other sources than patient interview, physical examination and patient chart	Ordering test, consulting colleague, seeking external information
Defining problem	Complex, interpretative assessments that define what the problem is and reflect a medically informed conclusion	Diagnostic conclusion, etiological inference, prognostic judgment
Treatment goal	Decision to set a defined goal for treatment and thereby being more specific than giving advice	Quantitative or qualitative
Drug	Decision to start, refrain from, stop, alter or maintain a drug regimen	Start, stop, alter, maintain, refrain
Therapeutic procedure	Decision to intervene on a medical problem, plan, perform or refrain from therapeutic procedures	Start, stop, alter, maintain, refrain
Evaluating test result	Simple, normative assessments of clinical findings and tests	Positive, negative, ambiguous test results
Deferment	Decision to actively delay a decision or rejection to decide on a problem presented by a patient	Transfer responsibility, wait and see, change subject
Advice and precaution	Decision to give the patient advice or precaution, transferring responsibility for action to the patient	Advice or precaution
Legal/insurance related	Medical decision concerning to legal regulations or financial arrangements	Sick leave, drug refund, insurance, disability

Table 1: Overview of the ten medical decision categories defined in MedDec, together with a brief description of each category and representative examples. The Table is reproduced from (Elgaar et al., 2024).

Decision Type	**Sex**		**Race**						**Lng. Proficiency**	
	Male (n=259)	**Female** (n=192)	**White** (n=327)	**AA** (n=42)	**Hispanic** (n=25)	**Asian** (n=15)	**NH** (n=1)	**Other** (n=21)	**En** (n=260)	**Non-En** (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 2: Percentage of annotated spans for each decision category across protected variables in MedDec. n is the number of the discharge summaries for each category. The last row shows the total count of decisions per variable.

2 MedExACT 2026 Task Description

The task is built on MedDec (Elgaar et al., 2024), which contains 451 discharge summaries, more than 54k sentences, and roughly 1.4M tokens annotated by domain experts. It reports token-level inter-annotator agreement of 0.74 Cohen's κ. The notes come from MIMIC-III, a de-identified critical care database distributed under controlled access (Johnson et al., 2016). MedExACT inherits that access model: participants obtain the training and validation data from PhysioNet, while the official gold labels for test set remain hidden.

Given a full discharge summary, systems must detect contiguous text spans that express medical decisions and assign each span one of nine DICTUM decision categories defined in Table 1: *Contact related, Gathering information, Defining problem, Treatment goal, Drug, Therapeutic procedure, Evaluating test result, Deferment, Advice and precaution*, or *None* for when no decision is present. Table 2 shows the distribution of the decision categories across protected variables in MedDec.

For MedExACT 2026, the released split contains 350 training summaries, 53 validation summaries, and the 48 hidden test summaries (Table 3). In addition to text and span annotations, the task package includes demographic attributes used for subgroup evaluation and phenotype metadata inherited from the MedDec release. For privacy reasons, access to the dataset requires approval for MIMIC-III under their data use agreement.

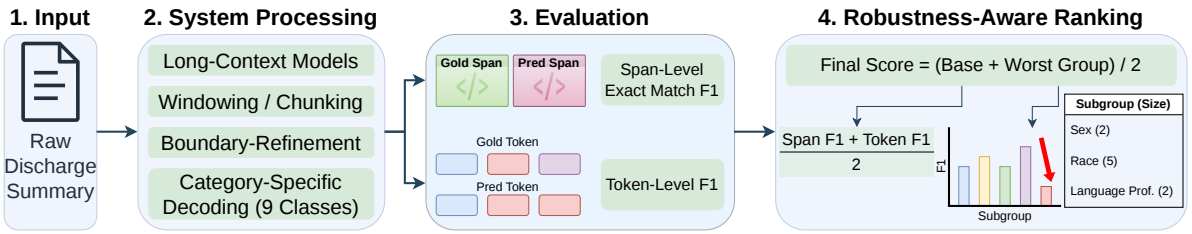

Figure 1: Overview of MedExACT 2026. (1) A raw ICU discharge summary from MIMIC-III is passed to the system. (2) Participating systems typically combine long-context encoders, windowing or chunking of long notes, boundary refinement, and category-specific decoding over the nine DICTUM decision classes to emit character-level span predictions. (3) Predictions are matched against hidden gold annotations at two granularities, span-level exact-match F1 (strict offset and label match) and token-level macro-F1, which are averaged into a per-note Base score. (4) The final leaderboard score is the mean of the overall Base score and the minimum Base score across nine subgroups defined by sex (2), race (5), and English proficiency (2), rewarding both accuracy and robustness.

As Figure 1 illustrates, the input to a submitted system is the raw text of a discharge summary taken from MedDec, and the output is a list of predictions, each represented by a start character offset, an end character offset, and a category ID. These predictions are matched against hidden gold annotations using a strict span-level evaluator together with a token-level evaluator, and the resulting Base score is combined with a worst-subgroup term to produce the final leaderboard score (Section 3). The shared task scores nine in-scope decision types. The underlying MedDec annotations follow DICTUM (Ofstad et al., 2016). Category 10 (Legal/insurance related) is ignored by the official shared-task evaluator.

Example: Each MedDec annotation includes annotator ID, discharge summary ID, and a list of decision spans with their text, DICTUM category, character offsets, and annotation ID:


```
{
 "annotator_id": "CALIML",
 "id": "10814_101543_52781",
 "annotations": [
  {
   "decision": "arrest with heart block",
   "category": "3: Defining problem",
   "start_offset": "526",
   "end_offset": "556",
   "annotation_id": "AC_005"
  },
  ...
  {
   "decision": "GU: Foley in place",
   "category": "6: Therapeutic procedure",
   "start_offset": "4000",
   "end_offset": "4018",
   "annotation_id": "AC_048"
  }
 ]
}
```


Split	#Notes	#Decisions	Access
Train	350	43,640	Released
Validation	53	7,044	Released
Test	48	6,537	Hidden

Table 3: MedExACT 2026 data split summary. The hidden test set is used only for official evaluation.

3 Evaluation

Submissions are assessed based on both performance and robustness across patient subgroups. Evaluation script is available on GitHub.[1]

Base Performance: is accessed by evaluating two core metrics: (i) *Span-F1*, which requires predicted spans to exactly match the gold spans (in both offset and label). The script performs a light word-boundary expansion and punctuation trimming before matching. (ii) *Token-F1*, computes macro-F1 per note after tokenizing the raw text with NLTK and then averages across notes. At token level, the evaluator excludes tokens that are covered by more than one gold span.

$$\text{Base} = (\text{SpanF1} + \text{TokenF1})/2$$

Robustness: Let $\mathcal{G}$ denote the set of subgroup buckets defined over sex, race, and language proficiency. The subgroup definitions used are Female/Male for sex, White/African American/Hispanic/Asian/Other for race, and English/Non-English for language. We compute the same Performance Base score separately for each of these nine subgroups and then calculate the minimum subgroup Base Score across all the

[1] github.com/CLU-UML/MedDec/blob/main/evaluate.py

Team	LLM	Ens	Cat	Long	Aug	Fair	Final
CUAMC (Baumgartner and Schilling, 2026)		✓				✓	0.596
LAMAR (Chiewhawan et al., 2026)	✓	✓			✓		0.594
Otter (Lowphansirikul and Ittichaiwong, 2026)		✓	✓		✓	✓	0.581
ELiRF-UPV (Ahuir et al., 2026)	✓		✓				0.581
MedMBZ (Elshehaby et al., 2026)		✓	✓	✓		✓	0.572
CASPAR (Tao et al., 2026)				✓			0.567
Kondadadi (Kondadadi, 2026)		✓		✓	✓		0.555
avishek		✓				✓	0.537
Fuyou Mao				✓		✓	0.533
ccccgo		✓		✓		✓	0.533
Aurum (Kumari et al., 2026)				✓	✓		0.525
dipika_nath		✓		✓	✓		0.514
TamuNLP		✓					0.512
Baseline				✓			0.511
CanSA (Alliheedi et al., 2026)	✓			✓			0.405
NoviceTrio	✓			✓			0.378

Table 4: Method matrix for the 15 system descriptions, sorted by descending final score, with the organizer Baseline (Elgaar et al., 2024) shaded for reference. LLM: LLM used in the prediction pipeline. ENS: multi-model ensemble or aggregation. CAT: category-specific decoding, routing, or reclassification. LONG: explicit long-document windowing or chunking. AUG: synthetic, pseudo-labeled, or augmented training data. FAIR: fairness-aware training, sampling, or model selection. Nearly all systems also use an encoder backbone and some form of boundary post-processing; those near-universal features are omitted.

subgroups. This metric rewards systems that optimize across all target subpopulations rather than optimizing only the global average; this explicitly incentivizes fairness-aware system design instead of treating fairness as a post-hoc diagnostic:

$$\text{WorstGroup} = \min_{g \in \mathcal{G}} \text{Base}_g$$

The final leaderboard ranking uses the following fairness-aware aggregate score:

$$\text{FinalScore} = (\text{Base} + \text{WorstGroup})/2$$

We consider the following two implementation details while evaluating: first, predictions must use character offsets into the exact raw text supplied to the evaluator. Any normalization of the whitespace, punctuation, or section structure can shift offsets and reduce scores; second, the shared task only scores category IDs 1–9. Gold spans or predictions outside that label space are ignored.

4 Task Resources and Public Baseline

In addition to the submission endpoint, the released repository includes official split files without text spans (the actual spans can be obtained from PhysioNet), a sample predictions JSON file, a data-cleaning utility for offset normalization, a statistics generator for subgroup evaluation, raw-text extraction scripts for MIMIC-III notes, and the official evaluator. The released bundle is also intended to support replication after the workshop period.

The public baseline fine-tunes RoBERTa for span tagging: tokens receive multiclass BIO-style labels over 512-token segments, and the resulting token labels are converted back to character-offset predictions for evaluation. The implementation also includes optional CRF decoding and random cropping of long notes during training. This baseline is meant to help participants verify preprocessing, training, and submission formatting before developing their own systems; it should not be treated as the definitive performance reference for the leaderboard. The original MedDec paper provides further baseline comparisons (Elgaar et al., 2024).

5 Participating Systems

The system-description file available to us contains 16 submissions from 15 organizations, and the leaderboard snapshot contains 37 scored runs (including the baseline model). Because the leaderboard file records submission usernames rather than canonical team names, our quantitative analy-

sis is run-centric, whereas the method analysis below uses the author-provided system descriptions when they are available.

Table 4 summarizes six design dimensions across the 15 described systems. Ensembles appear in 9 descriptions, explicit long-document windowing in 8, fairness-aware training or model selection in 6, data augmentation in 5, LLMs use in 4, and category-specific decoding or routing in 3. Nearly all systems also use an encoder backbone and some form of boundary post-processing; those near-universal features are omitted from the matrix.

The most discriminating column is category-specific processing. All three systems that route predictions through per-category decoders, reclassifiers, or routing branches—Otter (Lowphansirikul and Ittichaiwong, 2026) (category reclassifier), ELiRF-UPV (Ahuir et al., 2026) (independent CRFs per category with a section-category layer mixer), and MedMBZ (Elshehaby et al., 2026) (sparse category routing with span-length calibration)—were in the top five. This pattern suggests that a single set of decoding thresholds fits the nine heterogeneous DICTUM categories poorly and that category-aware specialization provides consistent gains. Ensembles and fairness-aware optimization also skew toward the top: four of the five highest-scoring systems use multi-model aggregation, and three explicitly optimize for the worst-group term in the official metric.

Explicit long-document windowing is common but does not predict high final scores on its own. The top two systems (CUAMC (Baumgartner and Schilling, 2026) and LAMAR (Chiewhawan et al., 2026)) do not flag windowing as a primary design choice, relying instead on architectures that handle length natively or on section-level decomposition. Among the four LLM-based systems, fine-tuned LLMs with ensemble aggregation (LAMAR) reach the top of the leaderboard, while prompt-based LLM pipelines (CanSA (Alliheedi et al., 2026), NoviceTrio) fall well below the median, confirming that supervision and structured decoding remain critical even when the base model is large.

6 Results

Table 5 reports the full test-phase leaderboard snapshot with all 37 submitted runs, using the organizer-provided display names listed in the paper. The top five entries are CUAMC, LAMAR, Otter, ELiRF-UPV, and MedMBZ. Across all 37 runs, the median final score is 0.533, while the mean of the top five final scores is 0.585. The organizers' baseline, shown as MedExACT Organizers, ranks 28th with a final score of 0.511. The gap between the median and the top runs shows that small modeling choices matter under this setup.

Submission	Final	Base	Worst	Span	Token
CUAMC	0.596	0.604	0.589	0.542	0.667
LAMAR	0.594	0.600	0.588	0.526	0.675
Otter	0.581	0.592	0.569	0.518	0.667
ELiRF-UPV	0.581	0.589	0.572	0.524	0.654
MedMBZ	0.572	0.585	0.560	0.490	0.680
NA	0.572	0.584	0.560	0.489	0.679
venus	0.572	0.585	0.558	0.490	0.680
fauna_rhea	0.571	0.584	0.558	0.490	0.678
jordanaskanov	0.570	0.582	0.557	0.489	0.675
CASPAR	0.567	0.573	0.560	0.506	0.641
NA	0.564	0.576	0.552	0.488	0.664
Kondadadi	0.555	0.577	0.534	0.482	0.673
fnkll	0.554	0.575	0.532	0.491	0.659
kotymoty	0.553	0.573	0.532	0.490	0.655
avishek	0.537	0.557	0.517	0.485	0.629
uijdsadada	0.533	0.541	0.525	0.430	0.652
ccccgo	0.533	0.541	0.525	0.430	0.652
csu-Medical	0.533	0.541	0.525	0.430	0.652
wubeining123	0.533	0.541	0.525	0.430	0.652
tekak_xo	0.529	0.549	0.508	0.466	0.632
rs-submission	0.529	0.549	0.508	0.466	0.632
popopop	0.528	0.535	0.521	0.438	0.632
Aurum	0.525	0.536	0.514	0.441	0.631
dbasudbasidad	0.515	0.523	0.508	0.412	0.633
dipika_nath	0.514	0.537	0.491	0.422	0.652
suxinqi	0.512	0.518	0.507	0.387	0.649
sususu123	0.512	0.517	0.506	0.385	0.649
Baseline	0.511	0.530	0.492	0.436	0.624
rm-rf-humans	0.510	0.544	0.476	0.461	0.627
nrgdedede	0.509	0.525	0.493	0.439	0.612
john_real	0.509	0.536	0.482	0.465	0.606
sususu1998	0.506	0.512	0.499	0.384	0.641
sususu666	0.477	0.483	0.470	0.349	0.617
CanSA	0.405	0.419	0.392	0.341	0.497
NoviceTrio	0.378	0.408	0.348	0.239	0.578
shahrin	0.264	0.274	0.255	0.215	0.334
mehulit22	0.031	0.047	0.015	0.006	0.088

Table 5: Full test-phase leaderboard snapshot with all submitted runs, sorted by descending final score of teams. The top five entries are marked with inline badge icons and subtle row shading.

The top systems outperform the baseline across almost all subgroups (Figure 2), however the strongest system is not uniformly best for every subgroup. The leaderboard also reveals a consistent trade-off between token overlap and exact span recovery (Figure 3). Among the five best final-score runs, CUAMC has the strongest span F1, while MedMBZ has the strongest token F1. Similar behavior appears further down the leaderboard: several runs reach token F1 in the 0.67–0.68 range while remaining below 0.53 final score because exact

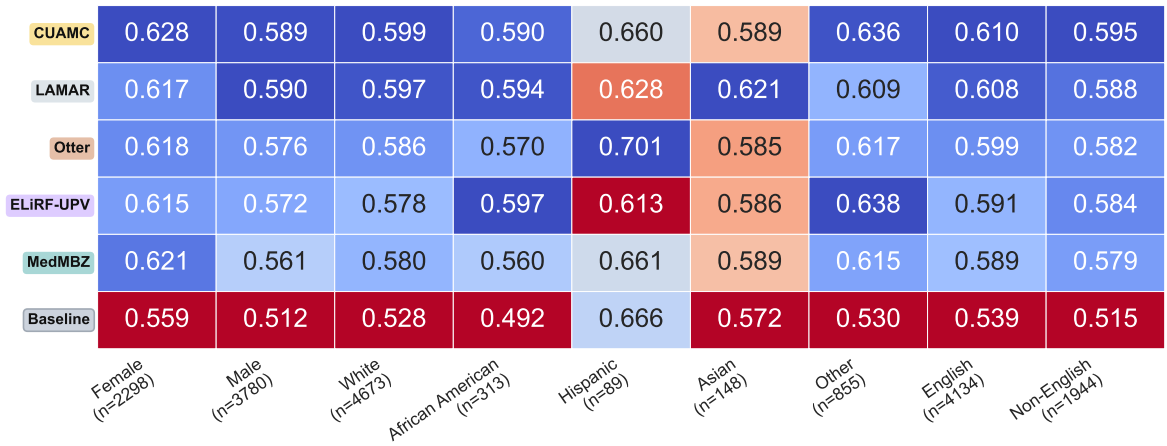

Figure 2: Subgroup Base scores for the top five MedExACT 2026 systems and the organizer baseline. Cell values show raw subgroup Base scores; colors are normalized within each subgroup to emphasize relative robustness patterns across sex, race, and English-proficiency groups.

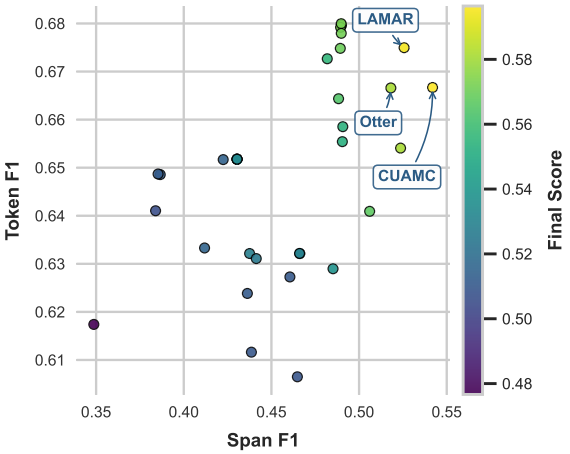

Figure 3: Scatter plot of Span F1 versus Token F1 across MedExACT 2026 leaderboard submissions. The top three systems (`CUAMC`, `LAMAR`, and `Otter`) are highlighted. High token-level overlap frequently fails to yield exact span boundary recovery.

span boundaries and subgroup robustness remain limiting factors. High token F1 with lower span F1 is consistent with coarse localization followed by boundary errors under the strict span matcher.

The fairness-aware objective also changes the interpretation of the ranking. For the top four runs, the worst-group score trails the base score by roughly 0.012–0.023 absolute points. These gaps are small enough to indicate deliberate robustness tuning, yet large enough to affect final ordering when systems optimize average accuracy only. Participant descriptions confirm that many teams treated the official final score as the primary development objective rather than as a secondary reporting metric.

7 Discussion

We summarize three takeaways from the participant reports and leaderboard snapshot:

Long-Context Modeling is a Core Requirement: Handling long discharge summaries is a prerequisite for this task. Given the length and density of ICU discharge summaries, most competitive systems used overlapping windows, chunk-level processing, or section-aware decomposition, including large encoder ensembles, Clinical-Longformer fine-tuning, and retrieval-based prompting pipelines (Kondadadi, 2026; Kumari et al., 2026; Alliheedi et al., 2026).

Exact Span Recovery is the Primary Bottleneck: Strong token F1 values show that systems can often detect the right region of text and recover the correct semantic type. The harder step is to return the precise character span expected by the official evaluator. Boundary refinement, CRF decoding, overlap-aware aggregation, and second-pass calibration all target this failure mode, so similar modules appear across otherwise different model families (Tao et al., 2026; Chiewhawan et al., 2026; Elshehaby et al., 2026).

Robustness-Aware System Design: Because the final ranking combines average performance with worst-group performance, teams invested in oversampling, subgroup-aware loss design, adversarial training, and development-time checkpoint selection under the official score (Baumgartner and Schilling, 2026; Lowphansirikul and Ittichaiwong, 2026; Elshehaby et al., 2026). For future clinically

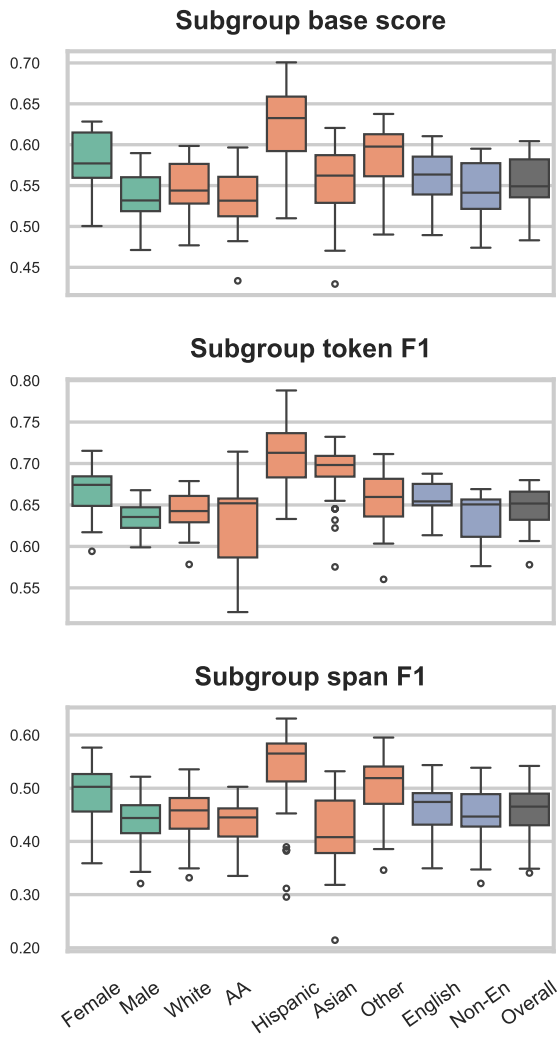

Figure 4: Distribution of Base score, Token F1, and Span F1 across submissions by subgroup and overall performance. Boxplots summarize variation across runs for sex, race, language, and overall scores, showing that subgroup differences persist across all categories.

grounded shared tasks, this example suggests that subgroup robustness should be built into the primary ranking rule instead of a post hoc diagnostic.

The Role of LLMs in Decision Extraction: The submitted systems also clarify the current role of LLMs in this space. Prior work has shown that large language models can be effective few-shot clinical information extractors in some settings (Agrawal et al., 2022). In MedExACT, however, long discharge summaries and strict character-offset evaluation seem to favor LLMs as augmentation, reranking, or repair modules instead of as standalone span extractors (Chiewhawan et al., 2026; Alliheedi et al., 2026). This matches the low zero-shot and one-shot exact-match results reported in the MedDec paper (Elgaar et al., 2024).

Toward Generalization: A natural next step for this shared task is to move beyond in-domain robustness and evaluate whether systems can generalize across clinically distinct settings under explicit domain shift. A future version of MedExACT could frame this as phenotype-family transfer, where models are trained without access to one target phenotype family and then tested on discharge summaries from that unseen family. This new setting measures whether systems learn decision semantics instead of phenotype-specific lexical or discourse patterns. We can annotate discharge summaries with their corresponding phenotypes, group phenotypes into clinically meaningful families, and reports leave-one-family-out results. We expect this setting to be feasible and result in substantially lower than in-domain performance due to the difficulty of transfer to unseen clinical contexts. This direction would complement the current robustness-aware evaluation by testing not only subgroup stability, but also resilience to shifts in patient populations and clinical conditions, and could motivate new methods for domain generalization, transfer learning, and clinically grounded robustness in medical decision extraction.

Understanding Clinical Decisions: This task can help us better understand underlying clinical decision-making by transforming unstructured clinical notes into structured representations of the decisions clinicians make. This enables large-scale analysis of how clinical reasoning is recorded across patients, conditions, and care settings. Such structured evidence could also provide better clinical decision support by making prior decisions easier to review, compare, and audit; helping clinicians recognize patterns in diagnosis and treatment; identifying inconsistencies in documentation (Amiri et al., 2024) and blind spots of extraction models (Elgaar and Amiri, 2026); and inform decision-support tools grounded in real clinical narratives.

8 Conclusion

MedExACT 2026 focuses on robust medical decision extraction over long discharge summaries. The submitted systems mainly use long-context encoder models, ensembles, boundary-aware decoding, and explicit optimization for the official worst-group-adjusted score. The task establishes a benchmark for medical decision extraction and shows how robustness-aware evaluation can influence system design in clinical NLP.

References

Micheal Abaho, Danushka Bollegala, Paula Williamson, and Susanna Dodd. 2021. Detect and classify – joint span detection and classification for health outcomes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8709–8721, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vicent Ahuir, Lluís Felip Hurtado, and María José Castro-Bleda. 2026. ELiRF-UPV@MedExACT 2026: Dynamic Section Conditioning for Medical Decision Span Detection in Discharge Summaries. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Mohammed Alliheedi, Robert E. Mercer, Anemily V. Machina, Sudipta Singha Roy, Yetian Wang, and Xindi Wang. 2026. CanSA at MedExACT@ACL 2026: Zero-Shot, Fine-Tuned, and Retrieval-Augmented Extraction of Clinical Decisions with Corpus Boundary Diagnostics. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Hadi Amiri, Nidhi Vakil, Mohamed Elgaar, Jiali Cheng, Mitra Mohtarami, Adrian Wong, Mehrnaz Sadrolashrafi, and Leo A Celi. 2024. Analysis of race, sex, and language proficiency disparities in documented medical decisions. *medRxiv*, pages 2024–07.

William Baumgartner and Lisa M. Schilling. 2026. CUAMC @ MedExACT 2026: Robust Ensemble Voting for Fair Medical Decision Extraction. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Monrada Chiewhawan, Keetawan Limaroon, and Titipat Achakulvisut. 2026. LAMAR at MedExACT 2026: Agreement-Driven Large Language Model Ensembles for Clinical Decision Extraction from Discharge Summaries. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Mohamed Elgaar and Hadi Amiri. 2026. Linguistic blind spots in clinical decision extraction. In *Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026)*, pages 46–54, Rabat, Morocco. Association for Computational Linguistics.

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.

Ahmed Elshehaby, Mohamed Abdalla, and Youssef MK Mohamed. 2026. Sparse Category Routing and Fairness-Aware Optimization for Medical Decision Extraction. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Alistair E. W. 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).

Rishik K. Kondadadi. 2026. Diverse Transformer Ensemble with Majority Voting for Medical Decision Extraction at MedExACT 2026. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Jyoti Kumari, Vinay Babu Ulli, and Anindita Mondal. 2026. Team Aurum at MedExACT 2026@ACL: Data Augmentation and Clinical Longformer Fine-Tuning for Medical Decision Extraction. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Shifeng Liu, Yifang Sun, Bing Li, Wei Wang, Florence T. Bourgeois, and Adam G. Dunn. 2021. Sent2Span: Span detection for PICO extraction in the biomedical text without span annotations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1705–1715, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lalita Lowphansirikul and Piyalitt Ittichaiwong. 2026. Otter at MedExAct2026: Diverse Encoder Ensemble for Medical Decision Span Detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 2023. DICE: Data-efficient clinical event extraction with generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15917, Toronto, Canada. Association for Computational Linguistics.

James Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer Seale, Jordan Swartz, Greg McKelvey, Hui Dai, Yi Yang, and David Sontag. 2021. CLIP: A dataset for extracting action items for physicians from hospital discharge notes. 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 1365–1378, Online. Association for Computational Linguistics.

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.

Jing Tao, Amir Eskandari, and Farhana Zulkernine. 2026. CASPAR: A Context-Aware Span Refinement Approach for Decision Support. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

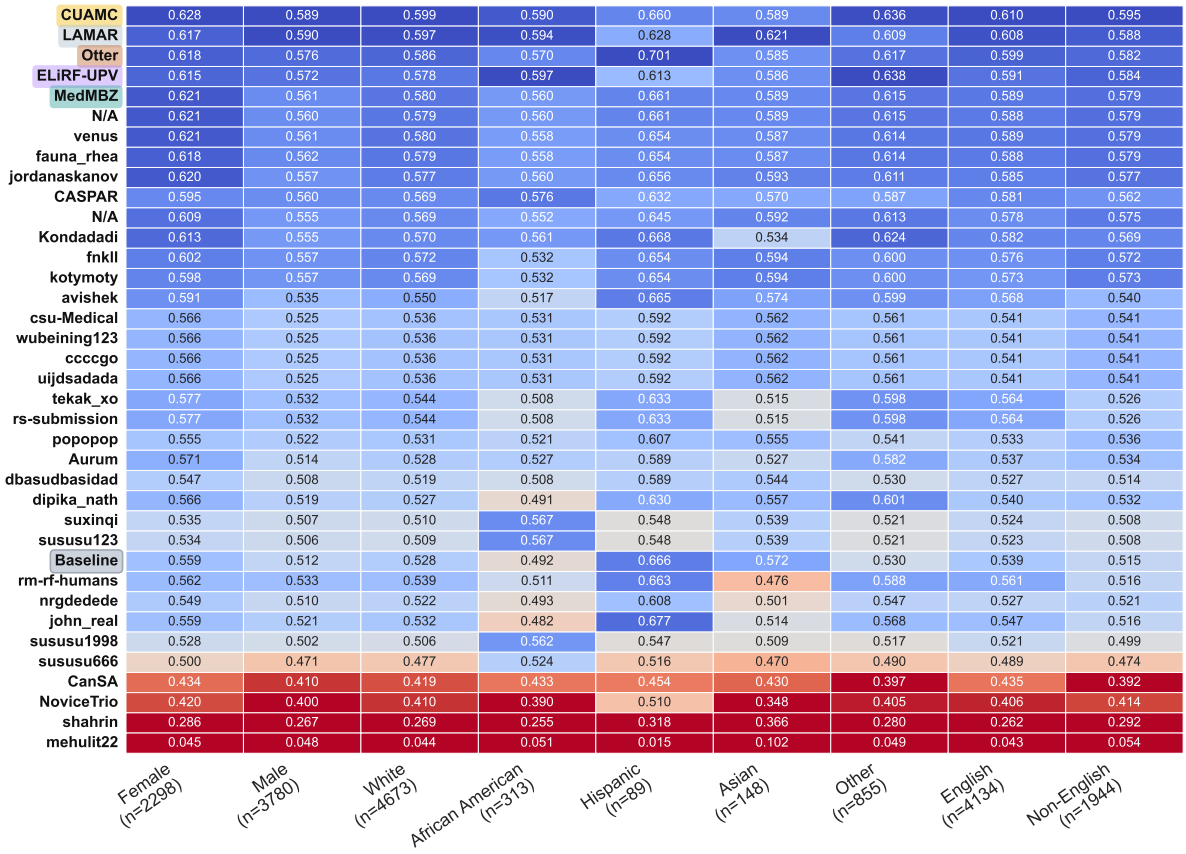

Figure 5: Full subgroup Base-score heatmap for all MedExACT 2026 leaderboard submissions. Rows are ordered by final leaderboard score; columns correspond to the nine sex, race, and English-proficiency subgroups used in the official robustness metric.

A Full Subgroup Leaderboard

Figure 5 shows the full subgroup Base-score heatmap for all MedExACT 2026 leaderboard submissions. Rows are ordered by final leaderboard score; columns correspond to the nine sex, race, and English-proficiency subgroups used in the official robustness metric. Three patterns stand out. First, the highest-ranked systems are strong across nearly all columns rather than winning through a single favorable subgroup; their Base scores mostly stay in the 0.58–0.66 range. Second, no system dominates every subgroup: `CUAMC` has the best English and Non-English scores, `LAMAR` is strongest on Asian patients, `Otter` is strongest on Hispanic patients, and `ELiRF-UPV` is strongest on African American and Other race groups. Third, subgroup robustness remains a real source of variation. Several mid-ranked runs have one or two noticeably weak columns, and the public baseline is illustrative: it is competitive on Hispanic patients but drops on African American, Male, and Non-English patients. The lowest-ranked systems are weak across all subgroups, which suggests that the robustness metric mainly separates otherwise capable systems rather than rescuing globally poor extractors.