

# CUAMC @ MedExACT 2026: Robust Ensemble Voting for Fair Medical Decision Extraction

**William A Baumgartner Jr and Lisa M Schilling**

Department of Medicine | Division of General Internal Medicine

University of Colorado Anschutz Medical Campus

Correspondence: [william.baumgartner@cuanschutz.edu](mailto:william.baumgartner@cuanschutz.edu)

## Abstract

Automated extraction of medical decisions from clinical notes is a critical step to constructing more granular patient health trajectories than what is currently obtainable from structured healthcare data. Here we present a system designed for the MedExACT shared task that employs an ensemble of BERT-based classifiers to account for demographic diversity when extracting mentions of medical decisions from MIMIC-III discharge summaries. A simple voting strategy combined with architectural diversity is demonstrated to work best when training data is limited.

## 1 Introduction

Structured clinical data detail the outcome of medical decisions by logging discrete observable events and state changes in a patient's health trajectory in the form of diagnoses, medications, laboratory measurements, procedures, and other data. These structured data, however, do not explicitly indicate the reasoning behind the medical decisions that led to the logged observable events. While simple reasoning for some medical decisions can be inferred from structured data (e.g., a new prescription following an abnormal lab measurement), intricacies such as patient-specific context (e.g., patient unable to tolerate side effect X), justifications based on differential diagnoses (e.g., likely Y, given Z, despite A), guideline-based decisions (e.g., skipping B because of C), and temporal aspects of decisions (e.g., pause in treatment due to D) are not explicitly captured. Clinical notes, however, often provide extensive information about medical decisions, including explicit statements about the reasoning for specific medical decisions. The ability to mine clinical notes for mentions of medical decisions represents the first step in automating the comprehensive understanding of a patient's health trajectory.

The MedDec corpus (Elgaar et al., 2024a,b) was designed to facilitate development and evaluation of automated systems for identifying mentions of medical decisions in clinical notes. MedDec supplements a subset of discharge summaries from the MIMIC-III corpus (Johnson et al., 2016b,a) with manually-defined span-level annotations labeled with ten medical decision categories from the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM) (Ofstad et al., 2016). The MedDec corpus defines "medical decision" broadly as "a particular course of clinically relevant actions and/or a statement concerning the assessment of a patient's health as defined in the DICTUM" (Elgaar et al., 2024a). This broad definition provides wide coverage for aspects of medical decision making, but also challenges automation of this extraction task due to the potential for overlapping categories (e.g., "there is no pericardial effusion" is annotated using "Defining problem" and "Evaluating test result", 76% and 24% of the time, respectively).

The Medical Decision Extraction, Analysis, and Classification Task (MedExACT) Shared Task @ ACL 2026 leveraged the MedDec corpus to prompt development of automated systems for identifying medical decisions from 9 of the 10 DICTUM categories in clinical text. The demographic makeup of the corpus and the intentional design of the evaluation metric also provides a test bed for studying generalization and robustness of extraction approaches across different demographic groupings. This manuscript describes the development of an automated approach to extracting mentions of medical decisions from clinical text using an ensemble of different classifiers to capture variance in the data and balance token-level and span-level performance generally while not penalizing specific demographic groups.

## 2 Methods

### 2.1 Data Preparation

Manual review of MedDec corpus annotations and documents resulted in the implementation of several data preprocessing steps. Analysis of the demographic groupings in the provided training/validation split of the data revealed a single Hispanic discharge summary in the validation set. Restratification of the data resulted in a more balanced distribution of demographic groups between the training and validation sets. Review of discharge summaries revealed two systematic formatting idiosyncrasies. First, 401 instances of presumably non-ASCII characters across 49 documents manifesting as a string of question marks ("?????") were discovered. These question mark strings were deduced to mainly be missing punctuation characters, and were fixed in a semi-automated way, while insertion of space characters maintained character offsets in the documents so as not to disturb the evaluation. Second, as it has been de-identified, the MIMIC-III corpus contains many instances of placeholders of previously identifiable information (e.g., Dr. [\*\*Last Name (STitle) \*\*]). Early experiments suggested that replacing the placeholders with synthetic data, e.g., "Dr. Smith", resulted in improved classification performance of medical decisions. These data preparation steps were used consistently for all model inputs.

### 2.2 Base Models

Preliminary experiments with several different models (ELECTRA-base (Clark et al., 2020), BioELECTRA (Kanakarajan et al., 2021), BiomedBERT (Gu et al., 2021), BioClinicalBERT (Alsentzer et al., 2019), and BioClinical-ModernBERT (Sounack et al., 2025)) suggested BiomedBERT was more amenable to the medical decision extraction task than other models, somewhat surprisingly even over models trained on the MIMIC-III corpus. The preprocessing steps to alter MIMIC-III artifacts likely impacted this result. Using BiomedBERT predominantly as a base, we developed an ensemble system comprised of six different architectural variants in an attempt to provide diverse responses to account for corpus variance (Table 1).

#### 2.2.1 Variant 1: BiomedBERT Plain

This variant is designed as a straightforward implementation of BiomedBERT that also serves as a

base model for other variants. It uses token-level BIO tagging (Ramshaw and Marcus, 1995) to encode the DICTUM annotations. Overlapping spans are excluded from the training and validation data inherently by the BIO representation. For the base variant, a 10-fold weighting of Hispanic documents via over-sampling was used due to the single Hispanic document in the original validation set.

#### 2.2.2 Variant 2: BiomedBERT + R-Drop

Variant 2 combines BiomedBERT with regularized dropout (Wu et al., 2021) (R-Drop;  $\alpha=3.0$ ) intended to reduce overfitting. It also incorporates section-to-DICTUM-category priors to leverage the fact that some categories are more likely to appear in certain note sections. The priors are applied to the B-tag logits at inference time to decrease the probability of rare category-section pairs, and increase the probability of frequent category-section pairs.

#### 2.2.3 Variant 3: BiomedBERT + DAPT

Domain-adaptive pre-training (DAPT) (Gururangan et al., 2020) was employed with BiomedBERT to construct Variant 3. BiomedBERT was further pre-trained using masked language modeling for 50k steps on the entire MIMIC-III corpus of discharge summaries (~59k notes). This pre-training step was entirely unsupervised and did not use any MedDec annotations or task-specific labels. As the MedDec dataset is derived from MIMIC-III, the pre-training corpus necessarily includes documents later used for downstream evaluation. This setup is consistent with common DAPT practice, where large unlabeled corpora are reused across pre-training and evaluation. Excluding test documents from DAPT, however, would provide a stricter separation and represents a potential refinement for future work. The resulting model checkpoint was also used as the base model for Variants 4 and 5.

#### 2.2.4 Variant 4: BiomedBERT + DAPT + Global Pointer

Variant 4 takes a span-based approach that is fundamentally different from the BIO representation used by the other five variants. Instead of labeling each token with B/I/O, and subsequently determining spans, the Global Pointer approach (Su et al., 2022) directly scores every possible (start, end) token pair for each DICTUM category. The primary motivation for including Variant 4 is to increase architectural diversity.

### 2.2.5 Variant 5: DAPT + R-Drop + Section Embeddings

Variant 5 builds on previous variants by combining DAPT with R-Drop and section priors, and adds section embeddings that are trained jointly with the model. The section embeddings encode which note sections (see Section 2.3) each token belongs to and are added to the token embeddings before the transformer layers. The intuition is that tokens in one section (e.g. "Discharge Medications") should be treated differently than tokens in a different section (e.g., "Assessment and Plan").

### 2.2.6 Variant 6: BioClinical-ModernBERT + R-Drop

This is the only variant not based on BiomedBERT. The BioClinical-ModernBERT architecture uses a different attention mechanism and is trained on a broader spectrum of biomedical and clinical data than BiomedBERT. The motivation is again to add diversity in responses as this model should make different errors than the other five variants. R-Drop is incorporated to minimize overfitting and section priors are also used.

## 2.3 Input Representation

Due to the limited context window for BERT-based models (512 tokens), discharge summaries must be split into chunks for processing. Document sections were iteratively derived from manual analysis of the training data, with canonical section names and their variations encoded as case-insensitive regular expressions, matched only on lines shorter than 80 characters to reduce false positives from narrative text. In total, 23 canonical document section types (Appendix Table 4) were defined and used to segment notes, enabling consistent alignment of semantically equivalent sections (e.g., "Discharge Medications", "Medications on Discharge") across documents. Text appearing prior to the first matched section header was labeled as "Preamble", and unmatched headers were treated as likely subsections and assigned to the most recent preceding canonical section.

Analysis of the training data showed that 99.7% of medical decision annotations fall within a single section, motivating a section-aligned chunking scheme in which input window boundaries respect section boundaries. Sections longer than 512 tokens are split into consecutive, non-overlapping sub-chunks. Short adjacent sections that fit entirely

in the 512 token window are merged into a single chunk.

## 2.4 Bias mitigation

Demographic subgroup bias mitigation focused on race/ethnicity subgroups based on early experiments. Three different demographic weighting schemes were used. Initial experiments, influenced by the single Hispanic document in the original validation set, applied 10-fold document-level oversampling of Hispanic documents (A in Table 1). One variant makes use of uniform weighting across all documents (B in Table 1), due to an unintended configuration override. Two variants weight the Hispanic, African American, and Other groups by 15x, 8x, and 6x, respectively (C in Table 1), designed to approximately equalize sampling rates across underrepresented groups. Sex and language proficiency were not explicitly modeled, however, the evaluation metric's explicit inclusion of worst-group performance guided model selection and threshold tuning across all experiments.

## 2.5 Ensemble composition

The full ensemble comprises 30 models: 5 random seeds for each of the 6 architectural variants. Each model processes every input chunk to classify mentions of DICTUM categories in the text. Candidate annotations are normalized using the `refine_span` function from the official evaluator code. Three different matching strategies are employed (Table 2). Run A employed a per-category voting scheme which accounts for some categories being easier to detect than others. It combines weighting for each variant with separate thresholds for each category, as well as pruning underperforming individual models. In the end, 26 of the 30 models are used, and the six variants are weighted 1/1/1/3/3/3 for a total of 56 votes. Category thresholds range from 8/56 for the rare Deferment category to 18/56 for the more frequent Drug category. Run B used a confidence-weighted voting scheme using the probability of the B-tag at the span's first token to scale each model's vote. Run B optimized for worst group score in an attempt to further mitigate demographic bias. Run C used a single global threshold with equal weights; at least 8 of 30 models must predict a span with the same category and normalized text. All voting thresholds were tuned on the re-stratified validation set.

#	Variant	Head	DAPT	R-Drop	Sec. Emb.	Sec. Prior	Wt. Grp.
1	BiomedBERT	BIO					A
2	+ R-Drop	BIO		✓		✓	B
3	+ DAPT	BIO	✓				A
4	+ DAPT (Glob.Ptr)	GP	✓				A
5	+ DAPT+RDrop+Sec	BIO	✓	✓	✓	✓	C
6	ModernBERT+RDrop	BIO		✓		✓	C

Table 1: Ensemble component architectures (5 seeds each, 30 models total). Head: BIO = token-level BIO tagging, GP = Global Pointer span extraction. DAPT = domain-adaptive pre-training on MIMIC-III. Sec. Emb. = learned section embedding. Sec. Prior = section-conditional logit adjustment at inference. Demographic oversampling weight groups: A = Hispanic 10 $\times$ ; B = uniform; C = Hispanic 15 $\times$ , African American 8 $\times$ , Other 6 $\times$ . ModernBERT = BioClinical-ModernBERT. All variants use section-aligned chunking with max length 512.

### 3 Results

Three ensemble systems comprising 5 randomly seeded models from each of the 6 architectural variants were constructed. Preliminary experiments showed that an ensemble of 5 seeds outperformed the base BiomedBERT model. This pattern was repeated for each of the other variants and amplified by combining the seeds from all six variants. The diversity of responses and errors made by the different seeds of the different variants contributed to more consistent performance across demographic groups. Table 2 summarizes the performance of the three systems against both the re-stratified validation set and the held-out test set. While aggregate metrics provide an overall view of performance, per-category results (Appendix Table 6) reveal substantial variation across DICTUM classes, with more frequent classes generally achieving higher performance than rarer classes. Table 3 details the per-demographic performance for the three systems. Of note, the simplest voting strategy (Run C) generalized best to the test set with a negligible difference in Final Score between the re-stratified validation and test sets, while the more sophisticated voting schemes of Runs A and B overfit to the validation set demographics.

Table 5 (Appendix) presents a quasi-ablation of the system components on the validation set. Among individual components, DAPT provided the largest improvement over the baseline (+0.024), followed by R-Drop (+0.014). Multi-seed ensembling of a single variant yielded modest gains relative to individual models (+0.027-0.029). The largest improvements came from combining architecturally diverse variants: the 15-model ensemble of Variants 1-3 reached 0.570, and each subsequent architecture increased overall performance incrementally, culminating in the 30-model ensemble at

0.597. Notably, the gains from architectural diversity (+0.045 for 15 models over the single-model baseline) substantially exceeded those from seed diversity alone (+0.027 for 5 seeds of one variant), suggesting that the ensemble benefits more from combining architectures that make different errors than from simply averaging multiple runs of the same model.

### 4 Discussion and Conclusion

Our ensemble approach to handling diversity in corpus annotation as well as in demographic variability proved to be effective at generalization in that it achieved the highest score on the test set. In practice, the simpler voting scheme of Run C was likely the difference maker as it did not experience noticeable overfitting to the validation data. Although Run B explicitly optimized for worst-group performance (primarily across race/ethnicity), this strategy did not generalize as effectively to the test set, suggesting that fairness-aware optimization may be sensitive to sampling variability in small datasets. Future work should explore more robust fairness-aware learning strategies, including explicit modeling of additional demographic attributes such as sex and language proficiency, subgroup calibration, and worst-group-aware training objectives, to better balance overall and subgroup performance. Ensemble diversity itself appears to improve worst-group robustness, even without extensive explicit subgroup modeling.

This work presents a robust baseline system that leverages conventional transformer-based approaches to identify mentions of medical decisions in text. Future work could involve several refinements and extensions to further improve robustness and generalizability. First, the heuristic voting strategy could be replaced with a learned span-

		Span F1	Token F1	WG	Final
<i>Restratified Validation (n=53)</i>					
Run A	26-model, per-cat thresh	.540	<b>.683</b>	.600	<b>.606</b>
Run B	30-model, conf-weighted	<b>.546</b>	.670	<b>.601</b>	.605
Run C	30-model, equal 8/30	.534	.677	.589	.597
<i>Test (n=48)</i>					
Run A	26-model, per-cat thresh	.547	.653	.547	.574
Run B	30-model, conf-weighted	<b>.553</b>	.654	.555	.579
Run C	30-model, equal 8/30	.542	<b>.667</b>	<b>.589</b>	<b>.597</b>

Table 2: Summary metrics for the three submitted runs. WG = Worst Group base score. Final Score =  $\frac{1}{2}$ Base +  $\frac{1}{2}$ [WG], where Base =  $\frac{1}{2}$ [Span F1] +  $\frac{1}{2}$ [Token F1]. Bold indicates the highest value in each column.

	Run A		Run B		Run C	
Subgroup	Val	Test	Val	Test	Val	Test
Female	<b>.600</b>	.614	.601	.618	.595	.628
Male	.619	.592	.612	.595	.612	.589
White	.610	.603	.605	.603	.606	.599
African Am.	.605	<b>.547</b>	<b>.601</b>	<b>.555</b>	.595	.590
Hispanic	.629	.648	.622	.635	.626	.660
Asian	.634	.607	.638	.600	.623	<b>.589</b>
Other	.602	.598	.603	.625	<b>.589</b>	.636
English	.614	.615	.609	.617	.612	.610
Non-English	.609	.577	.608	.582	.596	.595

Table 3: Base score ( $\frac{1}{2}$ [Span F1] +  $\frac{1}{2}$ [Token F1]) by demographic subgroup. **Bold** indicates the worst-performing subgroup for each run and split.

level fusion model to aggregate predictions across diverse ensemble members in a more principled manner, as explored in prior work on span-based system combination for named entity recognition (Fu et al., 2021). Second, fairness-aware learning could be strengthened through the strategies discussed above. Third, although domain-adaptive pre-training followed common practice, excluding evaluation documents from the pre-training corpus would provide a stricter separation and merits further investigation. Fourth, extending the extraction framework to natively support overlapping MedDec annotations may better capture the multi-label nature of clinical decision statements. Finally, deeper integration of document structure, such as through section-conditioned representations or decoding, may further exploit the strong concentration of specific medical decision categories within particular document sections observed in the data.

Although not discussed above, the exact span requirement in the evaluation metric proved challenging, especially considering some of the annotation inconsistencies (e.g., some annotations in a list include the bullet or number, and some do not) observed in the corpus. Significant effort went into analyzing and attempting to ameliorate such in-

consistencies, though none of the approaches were ultimately used. Restratifying the provided training and validation sets proved to be critical for developing models that generalize. More broadly, our results suggest that architectural diversity within an ensemble can compensate for limited and variable training data, and that restraint in threshold tuning is essential when validation sets are small.

## Acknowledgments

We thank the MedExACT shared task organizers and the MedDec corpus architects for providing the data and evaluation infrastructure, as well as the reviewers for their insightful comments. The MIMIC-III database was made available by PhysioNet. (Goldberger et al., 2000) AI-assisted coding tools (Anthropic Claude) were used during system development.

## Ethical Considerations

For any classification task, it is critical that the approach is designed to generalize over different population subgroups, especially those groups traditionally under-represented in medical studies. The MedExACT evaluation metric explicitly penalizes systems that perform poorly on specific

demographic subgroups. Our system design addressed demographic subgroups primarily through weighted sampling of documents from low frequency demographic subgroups.

MIMIC-III data was stored and processed exclusively on a HIPAA-compliant server managed by University of Colorado Anschutz Medical Campus Office of Information Technology. System development using AI coding assistance was conducted on a local machine without access to the clinical data; code was transferred to the server via a private repository.

## Limitations

Due to time constraints of the shared task, system development was driven largely by iterative experimentation rather than systematic ablation and tuning of different components. Once the ensemble approach proved effective for a single model (i.e., 5 random seeds improved over the single initial seed), development focused on expanding architectural diversity to handle annotation variance in the MedDec corpus, rather than optimizing individual model configurations. In the submitted runs, some leave-one-out experiments were performed to determine if removal of specific model seeds would improve overall performance (Run A), but this run overfit to the validation demographics. The best performing run (Run C) resulted from monotonic additions to the system that did not degrade performance on the "Worst-Group Base" score when evaluated on the restratified validation set. A prime avenue for future work would involve systematically ablating the contribution of each architectural component and optimizing the composition of the ensemble itself.

Despite the range of demographic categories captured by the MedDec corpus, the comparatively few examples of many of the categories, specifically the race/ethnicity categories, limit the ability to train models that will truly generalize well over a larger sample. This may have played a role in why the worst-performing demographic subgroup differed between validation and test for two of our three submissions (Table 3).

Finally, our use of 30 models in an attempt to capture annotation variance in the corpus, though it did achieve the highest score of all submitted results, is likely not a practical solution for production deployment. Processing each note 30 different times across an entire health data warehouse is

likely too inefficient without further optimization, such as pruning of redundant ensemble members.

## 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.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024a. [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, Hadi Amiri, and Leo Anthony Celi. 2024b. [MedDec: Medical Decisions for Discharge Summaries in the MIMIC-III Database](#). *PhysioNet*. Version 1.0.0.
- Jinlan Fu, Xuan-Jing Huang, and Pengfei Liu. 2021. Spanner: Named entity re-/recognition as span prediction. 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 7183–7195.
- Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8342–8360.
- Alistair Johnson, Tom Pollard, and Roger Mark. 2016a. [MIMIC-III Clinical Database](#). *PhysioNet*. Version 1.4.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016b. MIMIC-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: pre-trained biomedical text encoder using discriminators. In *Proceedings of the 20th workshop on biomedical language processing*, pages 143–154.

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*.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J Pollard, Eric Lehman, Alistair EW Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. *arXiv preprint arXiv:2506.10896*.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. *arXiv preprint arXiv:2208.03054*.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, and 1 others. 2021. R-drop: Regularized dropout for neural networks. *Advances in neural information processing systems*, 34:10890–10905.

## Appendix

#	Section Type
1	Allergies
2	Assessment and Plan
3	Attending Note
4	Brief Hospital Course
5	Chief Complaint
6	Discharge Condition
7	Discharge Diagnosis
8	Discharge Disposition
9	Discharge Instructions
10	Discharge Medications
11	Family History
12	Follow-up
13	History of Present Illness
14	Imaging / Radiology
15	Major Surgical / Invasive Proc.
16	Medications on Admission
17	Past Medical History
18	Past Surgical History
19	Pertinent Results
20	Physical Exam
21	Procedures
22	Review of Systems
23	Social History
24	Other / Unknown

Table 4: The 23 canonical section types identified in MIMIC-III discharge summaries, used for section-aligned chunking and section-conditioned inference.

Configuration	$n$	Final	$\Delta$
<i>Single model (best seed)</i>			
BiomedBERT (V1)	1	.525	—
+ R-Drop (V2)	1	.539	+.014
+ DAPT (V3)	1	.549	+.024
<i>5-seed ensemble</i>			
V1 $\times$ 5 seeds	5	.552	+.027
V3 $\times$ 5 seeds	5	.554	+.029
<i>Multi-architecture ensemble</i>			
V1–3	15	.570	+.045
+ V4 (Global Ptr)	20	.585	+.060
+ V5 (Sec-aware)	25	.592	+.067
+ V6 (ModernBERT)	30	.597	+.072

Table 5: Quasi-ablation on the validation set showing the contribution of individual components (top), multi-seed ensembling (middle), and progressive addition of architectural variants (bottom).  $n$  = number of models. Final = Final Score.  $\Delta$  is relative to the single BiomedBERT baseline.

#	Category	Gold	Run A			Run B			Run C		
			P	R	F1	P	R	F1	P	R	F1
1	Contact	361	.311	.543	.396	.344	.529	.417	.388	.438	.411
2	Gathering info	32	.222	.062	.098	.273	.094	.140	.273	.094	.140
3	Defining problem	2304	.534	.618	.573	.533	.630	.577	.512	.614	.559
4	Treatment goal	20	.529	.450	.486	.533	.400	.457	.421	.400	.410
5	Drug	1536	.584	.677	.627	.575	.689	.627	.556	.691	.616
6	Therapeutic procedure	700	.409	.490	.446	.486	.457	.471	.393	.484	.434
7	Evaluating test result	867	.420	.512	.462	.402	.502	.447	.405	.491	.444
8	Deferment	7	.000	.000	.000	.091	.143	.111	.000	.000	.000
9	Advice/precaution	274	.441	.730	.549	.462	.719	.563	.544	.653	.594
All (micro)			.491	.600	.540	.500	.601	.546	.488	.588	.534

Table 6: Per-category Span precision, recall, and F1 on the validation set for each submitted run. Gold shows the number of gold-standard spans. Note the class imbalance as some medical decision categories are sparsely represented.