Diverse Transformer Ensemble with Majority Voting for Medical Decision Extraction at MedExACT 2026

Rishik Kondadadi
konda052@umn.edu

Abstract

This paper describes our system for the MedExACT 2026 shared task on extracting and classifying medical decisions from ICU discharge summaries. We frame the task as BIO token classification and train 25 diverse transformer models spanning 13 distinct architectures, including Longformer, DeBERTa, RoBERTa, BioBERT, SciBERT, and others. Each model is trained with category-aware oversampling, focal loss, and demographic-group-aware sampling to address class imbalance and promote fairness across patient subgroups. At inference time, we aggregate predictions via text-normalized majority voting, retaining spans agreed upon by at least 6 of 25 models. Our best submission achieves a final score of 0.5554 on the test set, demonstrating that a simple vote-based ensemble over architecturally diverse models outperforms more complex filtering approaches. We find that architectural diversity is a key driver of ensemble quality and that cross-validation is essential for reliable model selection on small clinical datasets.

1 Introduction

Clinical discharge summaries from intensive care units (ICUs) contain critical medical decisions that guide patient care transitions. Automatically extracting and categorizing these decisions can support clinical decision support, quality assurance, and retrospective analysis. The MedExACT 2026 shared task (Elgaar et al., 2026) challenges systems to identify decision spans in MIMIC-III (Johnson et al., 2016) discharge summaries and classify them into nine categories from the DICTUM taxonomy (Elgaar et al., 2024): (1) Contact related, (2) Gathering information, (3) Defining problem, (4) Treatment goal, (5) Drug, (6) Therapeutic procedure, (7) Evaluating test result, (8) Deferment, and (9) Advice and precaution.

A distinguishing feature of this shared task is its fairness-aware evaluation. The final score is computed as the average of a base score (mean of span-level and token-level F1) and the Worst Group Score across nine demographic subgroups defined by sex, race, and language proficiency. This formulation penalizes systems that perform well on average but poorly on underrepresented populations.

We approach the task using a diverse ensemble of 25 transformer-based token classifiers. Our key findings are: (1) simple majority voting over diverse architectures outperforms learned ensemble methods, (2) architectural diversity matters more than model count within a single architecture family, and (3) cross-validation provides reliable model selection signals on small clinical datasets, while held-out validation scores are prone to overfitting.

2 System Description

2.1 Task Formulation

We formulate the task as BIO sequence labeling with 19 labels: O (outside), B-k (beginning of category k), and I-k (inside of category k) for $k \in \{1, \ldots, 9\}$. Special tokens and padding positions are assigned the ignore index -100 during training.

2.2 Model Architectures

Our ensemble comprises 25 models spanning 13 distinct pretrained architectures, chosen to maximize prediction diversity:

- **Longformer variants (11 models):** We use Longformer-base (Beltagy et al., 2020) and clinical Longformer variants trained with different hyperparameter configurations (learning rates from 2×10^{-5} to 5×10^{-5}, training durations from 5 to 50 epochs, random seeds 42 and 7). Longformer's sliding-window attention natively handles the long discharge summaries in the dataset, where approximately 32% exceed 4096 tokens.

Proceedings of the BioNLP 2026 (Shared Tasks), pages 87–91

- **DeBERTa-v3 (3 models):** We train DeBERTa-v3-base and DeBERTa-v3-large (He et al., 2023) with differential learning rates (encoder: 2×10^{-5}, classifier head: 1×10^{-4}) and sliding-window inference (window size 512, stride 128).

- **Domain-specific models (6 models):** BioBERT (Lee et al., 2020), Bio_ClinicalBERT (Alsentzer et al., 2019), BiomedBERT-large (Gu et al., 2022), SciBERT (Beltagy et al., 2019), BlueBERT (Peng et al., 2019), and PubMedBERT (Gu et al., 2022). These models are pretrained on biomedical or clinical corpora and bring domain-specific vocabulary knowledge.

- **General-purpose models (5 models):** RoBERTa-large (Liu et al., 2019), ELECTRA-large (Clark et al., 2020), ALBERT-xxlarge (Lan et al., 2020), XLNet-large (Yang et al., 2019), and BERT-large (Devlin et al., 2019). These models provide complementary representations learned from general-domain text.

2.3 Training Procedure

All models are fine-tuned as token classifiers using the HuggingFace Transformers library (Wolf et al., 2020). We employ several techniques to handle the challenges of the dataset:

Sliding-Window Chunking. For models with limited context windows (512 tokens), we chunk documents using a sliding window with a stride of 128 tokens. During training, overlapping windows share labels derived from the same gold annotations. During inference, predictions from overlapping windows are merged by retaining the highest-confidence prediction for each token position. Longformer models use a window size of 4096 with stride 1024.

Focal Loss. We use focal loss (Lin et al., 2017) with $\gamma = 2.0$ to address the severe class imbalance in the dataset, where the O label dominates:

$$\mathcal{L}_{\text{focal}} = -(1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

where p_t is the predicted probability of the true class. Additionally, per-class weights computed via inverse square-root frequency provide further emphasis on rare BIO labels.

Category-Aware Oversampling. Categories 2 (Gathering information), 4 (Treatment goal), and 8 (Deferment) are underrepresented in the training data. We assign training chunks containing these categories a sampling weight of $10\times$ through weighted random sampling, ensuring the model sees these rare categories more frequently during training.

Group-Aware Sampling. To promote fairness across demographic subgroups, we compute per-sample weights inversely proportional to the frequency of each document's demographic group, defined by the combination of sex, race, and language. This ensures underrepresented demographic groups receive proportionally more training signal.

Hyperparameters. We train with AdamW optimization, a warmup ratio of 0.1, weight decay of 0.01, and learning rates in the range $[2 \times 10^{-5}, 5 \times 10^{-5}]$. Models are trained for 5–50 epochs depending on convergence behavior, with gradient accumulation to achieve effective batch sizes of 16–32. The best checkpoint for each model is selected based on the official shared-task final score computed on the validation set at each epoch.

2.4 Inference

At inference time, each model processes a document using overlapping sliding windows as described above. BIO label sequences are decoded into spans by grouping consecutive B and I tags of the same category. Near-identical spans (same category, overlapping character offsets) are merged. Span boundaries are then refined by expanding to full word boundaries and trimming leading/trailing punctuation using the NLTK TreebankWordTokenizer, which matches the preprocessing applied in the official evaluation script.

2.5 Ensemble via Majority Voting

Given predictions from all 25 models for each document, we aggregate them using text-normalized majority voting:

1. For each model's predictions on a document, we refine span boundaries and extract the lowercased span text.

2. Predictions across all models are grouped by the key (category, normalized text).

3. A span is retained in the final output only if it receives votes from at least τ models. We set

$\tau = 6$ (approximately 24% of models).

4. For each retained span, the character offsets are taken from the contributing model with the highest confidence score.

This approach is deliberately simple. We experimented with more sophisticated aggregation methods—per-category voting thresholds, confidence-weighted voting, IoU-based span matching, and a gradient-boosted machine (GBM) filter trained to classify ensemble candidates as true or false positives—but found that uniform majority voting consistently produced the best and most robust results on cross-validation.

3 Experimental Setup

3.1 Data

The MedExACT dataset (Elgaar et al., 2024) consists of discharge summaries from the MIMIC-III database with over 56,000 expert-labeled decision spans across nine DICTUM categories. We use the provided data splits, with approximately 200 documents for training and 53 for validation. Each document is linked to patient demographic metadata (sex, race, language proficiency) used for the fairness-aware evaluation.

3.2 Evaluation Metrics

Systems are evaluated on a composite final score:

$$\text{Final} = \frac{\text{Base} + \text{Worst Group Score}}{2} \qquad (2)$$

where Base $= (\text{Span\_F1} + \text{Token\_F1})/2$ and the Worst Group Score is the minimum base score across nine demographic subgroups (Female, Male, White, African American, Hispanic, Asian, Other, English, Non-English).

Span F1 matches predictions to gold annotations by comparing tuples of (category, refined lowercased text), where span boundaries are preprocessed by expanding to word boundaries and trimming punctuation. Token F1 assigns token-level labels based on span overlap, computes macro-averaged F1 per document across the nine categories, and then averages across all documents.

4 Results

4.1 Test Set Results

We submitted three system configurations to the shared task. Table 1 summarizes the test set results.

System	Final Score
Organizer baseline (RoBERTa-base)	0.5111
Run 1: GBM filter on 25-model ensemble	0.4968
Run 2: Hybrid voting + LLM augmentation	0.4800
Run 3: 25-model vote ≥ 6	**0.5554**

Table 1: Test set final scores for our three submitted runs and the organizer-provided RoBERTa-base baseline (Elgaar et al., 2026). The baseline achieves Span F1 of 0.4363, Token F1 of 0.6238, Base of 0.5301, and Worst Group Score of 0.4922.

Our best submission (Run 3) exceeds the organizer-provided RoBERTa-base baseline (Elgaar et al., 2026) by 4.4 points absolute (0.5554 vs. 0.5111) and ranked 12th out of 37 submitted systems on the official leaderboard. The simple voting ensemble (Run 3) outperforms the GBM filter (Run 1) by 5.9 points absolute. The GBM filter was trained on the 53-document validation set to classify ensemble candidate spans as true or false positives using features such as vote count, confidence scores, and span length. Despite achieving a high validation score of approximately 0.75, the GBM overfits to the small validation set, resulting in substantially lower test performance. The hybrid approach (Run 2) supplemented the ensemble predictions on minority-demographic documents with spans extracted by a large language model (LLM) in a few-shot setting. This performed worst because the LLM extractions produced systematic boundary mismatches with the gold annotation conventions, yielding a validation score of only 0.277.

4.2 Subgroup Analysis

Our best submission (Run 3) attains its lowest base score on the Asian subgroup, which is notably absent from both the training and validation data. The vote-based ensemble handles unseen demographics reasonably well because it contains no learned parameters that could encode demographic bias—it simply counts agreement among independently trained models on the span text.

4.3 Ablation: Ensemble Size and Diversity

We compare our 25-model diverse ensemble against a 16-model ensemble restricted to Longformer and DeBERTa architectures only. On 5-fold cross-validation over the training set, the 25-model ensemble achieves a final score of approximately 0.50 versus 0.47 for the 16-model ensemble, confirming that architectural diversity—not simply model count—drives ensemble quality. Models

pretrained on different corpora and using different architectures (e.g., BioBERT vs. ELECTRA vs. XLNet) produce complementary errors, improving the signal-to-noise ratio of majority voting.

4.4 Cross-Validation vs. Validation Overfit

A critical finding from our experiments is the importance of cross-validation for model selection on this dataset. The 53-document validation set is too small for reliable evaluation: the GBM filter achieved 0.75 on validation but only 0.50 on test. In contrast, 5-fold cross-validation on the training set predicted final scores of approximately 0.49–0.50, which closely matched the actual test performance of 0.50–0.56. We used cross-validation scores for all ensemble configuration decisions, including the voting threshold τ and model selection.

5 Analysis and Discussion

Why simple voting wins. Learned ensemble methods (GBM filters, per-category thresholds) have sufficient capacity to memorize patterns in the small validation set. Majority voting has a single hyperparameter (τ), which limits its ability to overfit. The text-normalization step (lowercasing, boundary refinement) is crucial: it allows models with different tokenizers to agree on the same span despite minor offset differences.

Architectural diversity. Different pretrained architectures encode different inductive biases. Longformer captures long-range dependencies across the full document; DeBERTa uses disentangled attention representations; domain-specific models like BioBERT and ClinicalBERT bring biomedical vocabulary knowledge. Their disagreements are informative: spans confirmed across diverse architectures are more likely to be correct than spans supported by multiple variants of the same architecture.

Fairness through simplicity. Our group-aware sampling during training and the parameter-free nature of the voting ensemble contribute to relatively uniform performance across demographic subgroups. Since the ensemble aggregation does not use any document-level or demographic features, it cannot learn to behave differently for different patient populations.

LLM extraction limitations. Our attempt to use a large language model for few-shot span extraction scored poorly due to systematic boundary mismatches with the gold annotations. The LLM tended to paraphrase or extend spans beyond what the annotation guidelines specify, producing text that did not match after boundary refinement. This suggests that LLM-based extraction requires careful calibration to match dataset-specific annotation conventions.

6 Conclusion

We presented a 25-model diverse transformer ensemble for the MedExACT 2026 shared task on medical decision extraction. Our system aggregates predictions from 13 distinct pretrained architectures using text-normalized majority voting, achieving a final score of 0.5554. The key takeaways are: (1) architectural diversity in the ensemble matters more than model count or sophisticated aggregation methods, (2) majority voting provides natural regularization against overfitting on small clinical datasets, and (3) cross-validation is essential for reliable model selection when the validation set is small.

Limitations

Our system has several limitations. First, the voting threshold $\tau = 6$ was tuned via cross-validation on approximately 200 training documents and may not generalize to other clinical text domains or dataset sizes. Second, our approach requires training and running inference with 25 separate models, which is computationally expensive. Third, performance on demographic subgroups absent from training (e.g., Asian) remains lower, suggesting that additional techniques may be needed for true zero-shot fairness. Finally, our BIO formulation cannot represent overlapping or nested decision spans, which may occur in clinical text.

Ethics Statement

This work uses the MIMIC-III database (Johnson et al., 2016), which contains de-identified electronic health records. All data access was conducted under an approved PhysioNet credentialing agreement. Our system processes only de-identified text and does not attempt to re-identify patients. The fairness-aware evaluation framework of MedExACT highlights an important direction for equitable clinical NLP systems.

References

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

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. 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.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

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):160035.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

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*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32.