Otter at MedExAct2026: Diverse Encoder Ensemble for Medical Decision Span Detection

Lalita Lowphansirikul
VISTEC
lalita.l_s22@vistec.ac.th

Piyalitt Ittichaiwong
Siriraj Hospital, Mahidol University
piyalitt.itt@mahidol.ac.th

Abstract

We build an ensemble of 10 transformer encoders for the MedExACT 2026 shared task on medical decision span detection. The ensemble is diversified along three training directions: encoder initialization (including domain-adaptive pre-training on clinical text), loss function, and data augmentation with LLM-generated synthetic notes and silver-labeled clinical documents. Greedy forward search selects the combination with the highest validation score. The system achieves a final score of 58.09 (Span F1 51.81, Token F1 66.66) on the test set. Analysis shows that each added model differs from existing members along at least one training direction, producing span predictions that allow majority voting to screen out individual errors.

1 Introduction

Clinical discharge summaries document medical decisions such as diagnoses, prescriptions, and treatment plans, yet these decisions are embedded in unstructured text, making systematic analysis difficult. The MedExACT 2026 shared task (Elgaar et al., 2026) formalizes this as a span detection problem over a 9-category subset of the DICTUM taxonomy (Ofstad et al., 2016), adapting the MedDec dataset (Elgaar et al., 2024). The evaluation uses a subgroup-robust fairness metric that averages overall and worst-group performance across patient demographics (sex, race, and language).

This task poses two key challenges: (1) clinical discharge summaries often exceed the 512-token context window of standard encoders, requiring documents to be split into overlapping segments whose predictions must be reconciled; and (2) only a few hundred annotated documents are available for training across 9 fine-grained categories, three of which (Gathering Information, Treatment Goal, and Deferment) together account for a small fraction of annotations.

Rather than optimizing a single model, we train multiple encoders with different backbones, loss functions, and training data, then aggregate their predictions through majority voting, retaining only spans on which multiple models agree.[1]

2 System Description

We train a candidate pool of models that vary along three directions and apply greedy forward selection to identify the best-performing subset.

2.1 Task Formulation

We formulate decision detection as token-level BIO (Begin, Inside, Outside) sequence labeling. Each token is classified by a pretrained encoder followed by a dropout layer and a linear projection with softmax.

2.2 Candidate Pool

We vary three directions of training to build a diverse candidate pool: encoder backbone, loss function, and training data augmentation.

Encoder initialization. We train from two base encoders: BiomedBERT-base (109M parameters) (Gu et al., 2022) and GatorTron-base (345M parameters) (Yang et al., 2022). We additionally introduce a clinical variant of each via domain-adaptive pre-training (DAPT; Gururangan et al., 2020) with entity-centric masking (Lin et al., 2021), continuing the masked language modeling (MLM) objective on 7,814 MIMIC-III discharge summaries (Johnson et al., 2016), excluding MedExACT training and validation documents. We first identify *entities* in the MIMIC-III text: token spans labelled as medical decisions by a majority of five preliminary BIO taggers fine-tuned on the MedExACT training data. At each training step, 15% of tokens are then masked, with 80% sampled from positions inside

[1] Code available at https://github.com/lalital/diverse-encoder-ensemble-medexact

Proceedings of the BioNLP 2026 (Shared Tasks), pages 47–53

these entities and the remaining 20% drawn at random. DAPT runs for 5,000 optimizer steps with learning rate 5×10^{-5}, batch size 16, and cosine decay with 300-step warmup, yielding the DAPT-BiomedBERT and DAPT-GatorTron variants in the candidate pool.

Loss function. Each encoder is trained with both cross-entropy and focal loss (γ=2; Lin et al., 2017), which down-weights well-classified tokens and focuses training on ambiguous examples.

Training. Each model is trained independently as a single-task BIO tagger. Non-DAPT variants undergo a single supervised NER fine-tuning stage. DAPT variants undergo two sequential stages: self-supervised MLM (DAPT) followed by supervised NER fine-tuning, rather than a joint objective. We use AdamW (weight decay 0.01, batch size 16) for 8,000 steps with 6% warmup, with learning rates 3×10^{-5} for BiomedBERT variants and 2×10^{-5} for GatorTron variants. The schedule is cosine decay, except for the DAPT-GatorTron variant which uses linear decay. At inference, predictions from all models are combined through whole-span majority voting, retaining only spans that multiple models agree on (§2.4).

Data augmentation. The training set is augmented with two sources, each with its own labelling pipeline.

- **Synthetic notes.** MedGemma-27B (Sellergren et al., 2025) generates clinical notes with inline span markers of the form `<d cat="N">text span</d>`; we parse these markers into BIO labels (104 and 152 documents from two prompt iterations; see Appendix A and Figure 2).
- **Silver-labeled MIMIC-III.** Discharge summaries are labelled by a five-model ensemble of BiomedBERT and GatorTron models fine-tuned on the gold training data (Span F1 50.3 on the validation set).

During training, weighted random sampling draws 30% of each batch from the augmented pool on average. We train each backbone in three settings: train split, train split with synthetic notes, and train split with silver-labeled data, making data augmentation one of the dimensions varied across the candidate pool.

2.3 Input Segmentation

Documents are segmented into overlapping windows of 512 tokens with a stride of 384 during training, yielding 128 tokens of overlap between consecutive windows. All windows are used as training examples, ensuring full document coverage regardless of length.

At test time, the stride is reduced to 256 tokens so that each token receives predictions from more windows. Overlapping regions are merged via center-priority weighting: for a window of length L=512, each token at position t receives weight $w(t) = \exp(-(\frac{t-L/2}{L/2+1})^2)$, assigning higher weight to tokens closer to the window center.

2.4 Ensemble Strategy

We select the final ensemble by greedy forward selection (Caruana et al., 2004) from the candidate pool. The search begins with the highest-scoring individual model and iteratively adds the candidate that yields the largest improvement in final score, with the global voting threshold k re-optimized at every step. Selection terminates when no remaining candidate improves the score. Predictions are aggregated via span-level majority voting: each model's predicted spans are normalized (expanded to full words, punctuation-stripped, and lowercased) to form (category, text) pairs per document, and a span is included if at least k models predict the same pair. The final configuration uses 10 models with a default threshold of k=3. This composition is the deterministic outcome of greedy selection, not a designed coverage constraint: 6 of 10 models use augmented data, all three encoder initializations are represented, and both loss functions are present. A separate exhaustive search over per-category thresholds raises k to 4 for Defining Problem and to 5 for Advice and Precaution, reducing false positives in these high-recall categories.

3 Experiments

3.1 Setup

The MedExACT dataset (Elgaar et al., 2026, 2024) consists of clinical discharge summaries annotated with medical decision spans across 9 DICTUM categories. We use the provided training and validation splits. For the test submission, each model's best checkpoint (selected on the validation set) is used directly to generate test predictions.

The official evaluation metric, final score, averages the overall base score (mean of Span F1 and Token F1) and the worst-group base score (lowest base score among demographic subgroups defined by race, sex, and language); see Elgaar et al. (2026)

Encoder	Loss	Data	Span F1	Token F1	Final Score†
BiomedBERT	CE	+ 104 synthetic	42.7	61.6	50.5
BiomedBERT	Focal	Train split only	41.4	59.8	49.2
BiomedBERT	Focal	+ 104 synthetic	41.8	62.2	51.7
GatorTron	CE	+ 152 synthetic	41.2	62.6	49.1
GatorTron	CE	+ Silver annotated	45.0	63.8	52.7
GatorTron	CE	Train split only	43.9	63.4	51.2
BiomedBERT	CE	+ Silver annotated	44.8	61.1	50.5
GatorTron	CE	+ 104 synthetic	42.4	63.8	52.5
BiomedBERT	CE	Train split only	45.3	62.7	51.8
DAPT-GatorTron	CE	Train split only	45.0	64.1	48.6
10-model ensemble			50.4	62.7	56.9

Table 1: Ensemble models in greedy selection order, with majority vote k=3. Selection halted when no remaining candidate improved the validation final score. †Subgroup-robust final score on corrected validation set. Data: 104/152 synthetic = MedGemma-27B documents from two generation rounds at 0.3 mixing ratio, silver annotated = silver-labeled MIMIC-III documents, train split only = no augmentation.

for the full specification.

During development, we select checkpoints by final score on the validation set using corrected annotations, where 28 mislabeled Treatment Goal spans in a single Hispanic-subgroup document are reclassified as Defining Problem (§4); test submissions are scored by the organizers against the held-out test set with the original annotations. All models use 512-token sequences and are trained on NVIDIA A100 GPUs.

3.2 Results

Our system achieves a final score of 58.09 on the test set (Span F1 51.81, Token F1 66.66). Greedy forward selection yields a 10-model ensemble (Table 1) covering all three training directions. Notably, DAPT-GatorTron has the lowest score (48.6) yet is selected, suggesting that the ensemble benefits from diverse span predictions, allowing majority voting to screen out individual errors.

Per-category breakdown. Token F1 exceeds Span F1 across all categories (Table 2): the ensemble locates decision-relevant tokens but misaligns span boundaries. Drug Related achieves 91.0 Token F1 but only 57.8 Span F1, reflecting consistent drug vocabulary with variable boundary conventions. Contact Related shows the widest gap (76.1 vs 36.8), likely because referral and follow-up spans vary in how much surrounding context is included. Rare categories (Gathering Information at 12.2/26.2, Deferment at 15.4/23.0) remain low on both metrics, consistent with their low frequency in the training data.

Category	Span F1	Token F1
Contact Related (CR)	36.8	76.1
Gathering Information (GI)	12.2	26.2
Defining Problem (DP)	55.7	77.4
Treatment Goal (TG)	40.0	57.4
Drug Related (Dr)	57.8	91.0
Therapeutic Procedure (TP)	39.3	63.7
Evaluating Test Result (ETR)	35.5	68.9
Deferment (De)	15.4	23.0
Advice and Precaution (A&P)	54.9	70.4

Table 2: Per-category Span F1 and Token F1 of the 10-model ensemble on the corrected validation set.

4 Analysis

Ensemble composition. Greedy selection produces an ensemble covering all three encoder initializations, both loss functions, and all data augmentation variants (Table 1). This coverage emerges from the selection criterion, not by design: each added model differs from existing members along at least one direction.

False negative patterns. Two-thirds of false negatives (1,936 of 2,905) are missed by every individual model in the ensemble (oracle-missed). Over half of these oracle-missed spans (54%) cross line boundaries, and long spans (11+ words) account for 34% of misses compared to 21% for short spans (1–3 words). Multi-line decision spans and long spans appear particularly difficult, even with overlapping sliding windows.

Category confusion. The dominant error pattern is bidirectional confusion between Defining Problem (DP) and Evaluating Test Result (ETR). For example (target spans underlined), "Head CT re-

Error type	Count	%
Which boundary is wrong		
Start only	1,513	62.0
End only	633	25.9
Both	294	12.0
Direction		
Under-extension (pred shorter)	1,487	60.9
Over-extension (pred longer)	953	39.1

Table 3: Boundary mismatch breakdown for the 10-model ensemble on the validation set (2,440 spans with $\geq$30% overlap and matching category).

vealed a SAH with IVH. Neurosurgery was consulted for further management." is labeled ETR in the annotations but predicted as DP. Conversely, "She was started on tamiflu...and has completed a 5 day course. BPs have been stable to slightly elevated...Pulmonary nodule stable." is labeled DP but predicted as ETR; the model likely confused a clinical observation for a test result. Both directions reflect the difficulty of distinguishing problem definitions from test findings in clinical text.

Span boundary mismatch. Of 2,440 predicted spans that overlap a gold span of the same category (character-level intersection-over-union $\geq$0.3) but differ in offsets, 62% have only the start boundary wrong, 26% only the end, and 12% both (Table 3). The ensemble exhibits a conservative bias: 61% of boundary errors are under-extensions where the predicted span is shorter than the gold span, versus 39% over-extensions. The missed tokens are often clinically meaningful. For under-extension, the model drops clinical indications, e.g., predicting "*Alprazolam... as needed for*" while the gold span extends to include "*Alprazolam... as needed for anxiety*"; the indication word anchors the prescribing rationale but falls outside the model's predicted boundary. For over-extension, the model absorbs related findings, e.g., predicting "*Mildly thickened mitral valve leaflets. Mild mitral annular calcification. No MS.*" while the gold span ends before the negative finding "*No MS.*"; the model treats the negative exclusion as part of the same evaluative statement.

Boundary error recoverability. Error analysis of the ensemble's false positives suggests that most are located adjacent to gold spans with incorrect offsets rather than in unrelated text. As an exploratory diagnostic, we trained a BiomedBERT-base boundary refiner on validation data. For each ensemble-predicted span, the refiner embeds it in a context window with special marker tokens indicating the predicted boundaries. For example, if the ensemble predicts "*hydrocortisone 100mg IV*" as a Drug span but the correct span is "*She was given hydrocortisone 100mg IV*", the refiner input is:

```
...was given [START] hydrocortisone 100mg
IV [END] and started...
```

The model predicts where the span should actually start and end, similar to extractive question answering; here it would move the start marker left to include "She was given". Training data is generated from the validation set by independently shifting the start and end offsets of each reference span by $\pm n$ words ($n \leq 5$); the refiner is trained to recover the original boundary (AdamW, lr 2×10^{-5}, batch size 32, 5 epochs).

On validation data, the refiner raises the final score from 56.90 to 73.79. Because training and evaluation draw from the same validation gold spans, this value is an in-distribution ceiling rather than a held-out generalization estimate. The refiner was applied to ensemble-predicted spans in our test submission.

Annotation quality. Worst-group analysis revealed that a single validation document accounts for the Hispanic subgroup's low score: 28 spans annotated as Treatment Goal appear to be Defining Problem (e.g., "Denies chest pain..." in Review of Systems, "Regular rate and rhythm" in Physical Exam). Because this subgroup contains only one validation document, the subgroup-robust metric is sensitive to individual annotation inconsistencies.

5 Conclusion

We presented a 10-model diverse encoder ensemble for the MedExACT 2026 shared task, scoring 58.09 on the test set. Greedy selection consistently prefers models that differ from existing members, and error analysis identifies two residual challenges: start-boundary errors (62% of boundary mismatches) and Defining Problem/Evaluating Test Result confusion, both of which may benefit from incorporating clinical domain knowledge, for example through expert-guided annotation refinement or models that explicitly represent clinical reasoning structure.

Limitations

The ensemble, synthetic data generation, and silver labeling are designed for this task. With only 53

validation documents, tuning decisions carry overfitting risk; the test score (58.09) exceeds the validation ensemble score (56.9), but worst-group estimates are noisy as some subgroups contain fewer than ten documents. Because test annotations are held out, we cannot ablate the boundary refiner; the alignment between test and pre-refiner validation scores is suggestive but not controlled.

The synthetic and silver-labeled documents were not validated by clinical experts. Their value is primarily distributional: they expose the encoder to clinical vocabulary patterns and decision-span boundary conventions, not to clinically accurate reasoning. Despite this, six of ten selected members use augmented data, indicating a net benefit for token-level span detection even when the underlying clinical content may be implausible.

Acknowledgements

We used computing resources from the LANTA HPC system at the National Science and Technology Development Agency (NSTDA), Thailand.

Ethics Statement

This work uses de-identified MIMIC-III data (Johnson et al., 2016) under the PhysioNet Data Use Agreement. We release correction scripts for the annotation inconsistencies identified in §4, noting these reflect our judgment and have not been verified by the shared task organizer. Synthetic clinical notes generated by MedGemma-27B may contain clinically implausible content and should not be used for medical decision-making.

References

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. 2004. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04. ACM.

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.

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, Online. 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*, 3:160035.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online. Association for Computational Linguistics.

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.

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.

Andrew Sellergren and 1 others. 2025. MedGemma technical report. *arXiv preprint arXiv:2507.05201*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194.

Iteration 1

Brief Hospital Course:
[Multiple units of packed red blood cells (pRBCs)]^{TP} and fresh frozen plasma were administered. [Norepinephrine infusion]^{Dr}, titrated to achieve [target mean arterial pressure (MAP) $>$ 65 mmHg]^{TG}. [A central venous catheter was placed]^{TP} for hemodynamic monitoring. Initial labs demonstrated [hemoglobin of 6.5 g/dL]^{ETR}, [platelets of 80,000/uL]^{ETR}, and [lactate of 4.2 mmol/L]^{ETR}. [An ABG was drawn]^{GI} revealing acute respiratory acidosis. [Bilateral pulmonary contusions and pneumothoraces]^{ETR} on repeat CXR. [Bilateral chest tubes were placed]^{TP} under image guidance. [Fentanyl 100 mcg/hr]^{Dr} and [Propofol 4 mg/kg/hr]^{Dr} were initiated for sedation. [Nephrotoxic antibiotics were deferred pending culture results]^{De}. [Blood cultures were obtained]^{GI} on admission. [Patient remained hemodynamically labile]^{DP} requiring ongoing vasopressor support. [Consultations were made with Trauma Surgery, Pulmonology, and Nephrology]^{CR}.

Discharge Medications:
1. [Oxycodone 5mg PO Q4-6H PRN pain]^{Dr}
2. [Omeprazole 20mg PO QD]^{Dr}

Discharge Instructions:
[Report any dark or black stools, vomiting blood, or abdominal pain immediately]^{AP}.

Iteration 2

Brief Hospital Course:
Upon arrival, the patient was hypotensive, tachycardic, and tachypneic. Initial labs revealed [significant hyperkalemia (7.5 mEq/L)]^{ETR}. The patient was started on [bicarbonate and insulin with glucose drip]^{Dr}. [Hemodialysis was initiated emergently]^{TP} due to [refractory hyperkalemia]^{DP}. [CXR demonstrated pulmonary edema]^{ETR} consistent with fluid overload. [Maintain adequate hydration to prevent dehydration]^{TG}. [Control blood sugar levels]^{TG}.

[Hyperkalemia]^{DP}. [Volume overload]^{DP}. [End-stage renal disease]^{DP}. [Anemia]^{DP}.

[Renal function labs to monitor kidney health]^{GI}. [Complete blood count (CBC) to monitor hemoglobin level]^{GI}. [Avoid strenuous activity until cleared by your physician]^{De}.

Discharge Medications:
1. [Lisinopril 20mg PO daily]^{Dr}
2. [Furosemide 40mg PO daily]^{Dr}

Discharge Instructions:
[Monitor blood pressure and heart rate closely]^{AP}. [Strictly adhere to dietary sodium restriction (<2 grams per day)]^{AP}.

Followup Instructions:
[Follow up with Nephrologist within 2 weeks]^{CR}.

Legend: **CR** Contact Related **GI** Gathering Info **DP** Defining Problem **TG** Treatment Goal **Dr** Drug Related **TP** Therapeutic Proc. **ETR** Eval. Test Result **De** Deferment **AP** Advice & Precaution

Figure 1: Annotated excerpts from synthetic discharge summaries generated by MedGemma-27B. Colored brackets denote decision spans; superscripts indicate DICTUM categories. Iteration 1 (left) produces longer narrative text with annotations embedded in clinical sentences; Iteration 2 (right) combines narrative with structured lists. Both documents cover all 9 categories.

A Synthetic Data Generation Details

A.1 Generation Pipeline

Synthetic notes are generated by MedGemma-27B (Sellergren et al., 2025) via vLLM (Kwon et al., 2023) on 4×A100 40GB GPUs (temperature 0.8, top-p 0.95, max tokens 4,096, repetition penalty 1.1); the prompt was refined using Claude Opus 4.6. Iteration 1 (104 docs) uses DICTUM definitions, a few-shot example, and per-document demographics; Iteration 2 (152 docs) revises category targets to reduce DP/Dr imbalance (Figure 1).

The prompt (Figure 2) conditions each document on a sampled age group, sex, ICU care unit (weighted toward MICU/CCU), and one of 25 diagnostic profiles (e.g., STEMI, ARDS, sepsis, TBI); demographics are oversampled for underrepresented subgroups. It enforces per-category minimums, boundary rules with correct/incorrect examples, contrastive category disambiguation, negative examples, span length targets, and a few-shot fragment covering all 9 categories.

System: You are a clinical documentation specialist generating synthetic discharge summaries for medical NLP research. . .

User:
Generate a complete ICU discharge summary with inline annotations marking EVERY medical decision.
Format: <d cat="N">text span</d>
Categories: 1=Contact, 2=Test ordered, 3=Diagnosis, 4=Treatment goal, 5=Drug, 6=Procedure, 7=Test result, 8=Deferment, 9=Advice

Patient: *{age}* *{sex}*, admitted to *{care_unit}* for *{principal_dx}*.
{secondary diagnoses and comorbidities}

Requirements:
- Include sections: Chief Complaint, HPI, Brief Hospital Course, Pertinent Results, Discharge Medications, Instructions, Followup
- Mark ALL decisions. Aim for 60–120 annotations.
- Category minimums: $\geq$8 Cat 3, $\geq$5 Cat 5, $\geq$5 Cat 6, $\geq$5 Cat 7, $\geq$3 Cat 2, $\geq$3 Cat 4, $\geq$2 Cat 8

{10 boundary rules, e.g.:
CORRECT: <d cat="5">She was given hydrocortisone 100mg IV</d>
WRONG: <d cat="5">hydrocortisone 100mg IV</d> *(missing verb)}*
{Cat 3 vs Cat 7 disambiguation, e.g.:
Cat 3: <d cat="3">bilateral pneumonia</d> *(diagnosis)*
Cat 7: <d cat="7">CXR showed bilateral infiltrates</d> *(test result)}*
{Negative examples, e.g.:
NOT tagged: "BP was 85/35 and drifted as low as 57/35" *(context, not a decision)}*
{Span length guidance, e.g.:
Short: <d cat="3">COPD</d> Medium: <d cat="1">transferred to [**Hospital**]</d>
Long: <d cat="8">Bronchoscopy was deferred given clinical improvement</d>
Target: 30% short, 40% medium, 20% long, 10% very long}
{Few-shot example fragment covering all 9 categories}

Figure 2: Prompt template for synthetic note generation. Italicized sections are condensed; the full prompt includes 10 boundary rules, category disambiguation pairs, and a few-shot example covering all 9 categories.