LAMAR at MedExACT 2026: Agreement-Driven Large Language Model Ensembles for Clinical Decision Extraction from Discharge Summaries

Monrada Chiewhawan[*1,2], **Keetawan Limaroon**[*3], **Titipat Achakulvisut**[1],
[1]Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom, Thailand,
[2]Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand,
[3]Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Bangkok, Thailand,
Correspondence: titipat.ach@mahidol.ac.th

Abstract

Clinical decision extraction from discharge summaries detects contiguous text spans expressing medical decisions and assigns each to predefined categories. In this paper, we propose an ensemble approach using large language models for clinical decision extraction from discharge summaries in the MedDec dataset with XML-like inline tag annotations. The ensemble consists of Qwen3.5-4B models trained under three different settings: (1) Dynamic Fine-tuning (DFT) with LoRA on the original training set, (2) DFT with LoRA then GRPO reinforcement on the original training set, and (3) DFT with LoRA on the original training set augmented with pseudo-labels. We aggregated predictions for each document by category using weights derived from inter-model agreement. Agreement-driven ensembles further enhanced performance across all metrics, yielding a 8.31% gain in Overall F_1 over the baseline and securing second place on the test leaderboard. Subgroup analysis further confirms that performance remains consistent across demographic groups, with no disproportionate degradation on underrepresented populations. We release our code at `https://github.com/biodatlab/medexact-lamar`.

1 Introduction

Clinical decision extraction focuses on retrieving relevant medical decisions from unstructured medical texts. Ofstad et al. (2016) categorized clinical decisions into ten categories, providing a standardized taxonomy for developing extraction techniques and models.

MedExACT is a shared task focused on medical decision extraction from ICU discharge summaries using the MedDec dataset (Elgaar et al., 2026, 2024). MedDec contains 451 discharge summaries from MIMIC-III (Johnson et al., 2016) annotated under the DICTUM guideline. However, its population skews White, male, and English-speaking, making consistent performance across demographic groups difficult to achieve.

Encoder-only transformer models are used as baselines for this task. Their small context windows and token-wise classification limit performance, as they struggle to capture context across documents. Despite RoBERTa (Liu et al., 2019) being the strongest baseline with a Base Score of 0.5301 and Overall F_1 of 0.5111, the relatively low Span F_1 (0.4363) indicates that accurate decision boundary extraction remains a challenge. Moreover, zero-shot LLaMA-3-8B-Instruct (Grattafiori et al., 2024) reported in MedDec showed limited effectiveness, likely due to challenges in handling long contexts and generating structured outputs. These results suggest considerable room for improvement and highlight the difficulty of the task.

We propose an ensemble approach for clinical decision extraction based on inline XML-like tagging. The ensemble consists of models trained under three complementary settings: (1) DFT with LoRA on the original training set, (2) DFT with LoRA then GRPO on the original training set, and (3) DFT with LoRA on the original training set augmented with pseudo-labels. Our system achieved a Span F_1 of 0.5257, a Token F_1 of 0.6750, and a Worst Group F_1 of 0.5881, yielding an Overall F_1 score of 0.5942 on the leaderboard.

2 Related Work

2.1 From Sequence Labeling to Generative Extraction

Encoder models have long been the backbone of named-entity recognition, with BERT (Devlin et al., 2019) and its variants such as RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and ALBERT (Lan et al., 2020), setting strong baselines through masked language modeling. However, these models share a fundamental limitation. For example,

[*]Equal contribution.

Proceedings of the BioNLP 2026 (Shared Tasks), pages 179–190
July 3-4, 2026

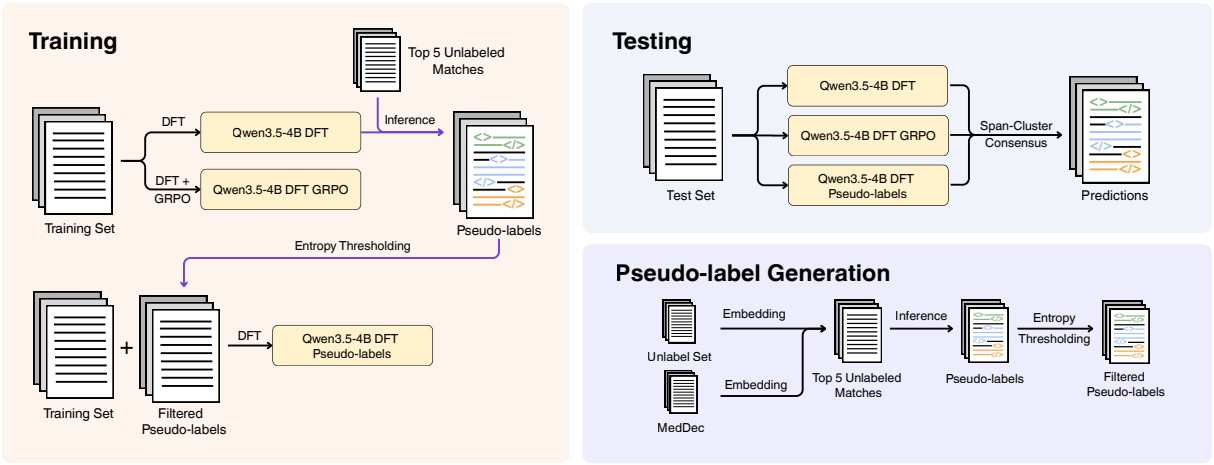

Figure 1: An overview of our system in training, testing, and pseudo-label generation. (Left) Training combines direct fine-tuning (DFT) and Group Relative Policy Optimization (GRPO) on Qwen3.5-4B, augmented with entropy-filtered pseudo-labels from unlabeled data. (Right) At test time, predictions from all three model variants are aggregated via Span-Cluster Consensus.

BIO tagging schemes assign exactly one label per token, making it impossible to represent overlapping spans regardless of model size or architecture. This is a critical shortcoming for clinical NER, given that 4.2% of labeled tokens in MedDec (Elgaar et al., 2024) belong to more than one span simultaneously.

Large language models offer an alternative approach. NER can be reframed as a generation task where the model rewrites the input with entity markers inserted inline, placing no structural restriction on overlapping spans. Wang et al. (2025) demonstrated this paradigm with GPT-NER using special token markup, achieving performance comparable to supervised baselines. Hu et al. (2024) showed that wrapping entities in HTML tags maps cleanly to standard evaluation schemes. However, with long, domain-specific, and complex documents like discharge summaries, we hypothesize that fine-tuning these models can serve as an effective approach for clinical decision extraction.

2.2 Supervised Fine-Tuning, Parameter-Efficient Adaptation, and Generalization

Supervised Fine-Tuning (SFT) trains LLMs on labeled examples to produce desired outputs. Since full fine-tuning is computationally expensive, Low-Rank Adaptation (LoRA) (Hu et al., 2022) and Weight-Decomposed Low-Rank Adaptation (DoRA) (Liu et al., 2024) offer a practical alternative by inserting small trainable matrices that capture task-specific changes with fewer parameters. However, SFT tends to memorize rather than generalize, where uncertain tokens receive disproportionately large gradients, pushing the model to overfit. Dynamic Fine-Tuning (DFT) (Wu et al., 2026) addresses this by scaling each token's loss according to model confidence, producing balanced updates that generalize substantially better across challenging tasks.

2.3 Reinforcement Learning with Verifiable Rewards

Reinforcement Learning with Verifiable Rewards (RLVR) replaces the learned reward model in classical RLHF (Ouyang et al., 2022) with verifiable reward functions. Group Relative Policy Optimization (GRPO) (Shao et al., 2024), improving on PPO (Schulman et al., 2017), derives training signals by comparing outputs within a sampled group using rule-based rewards, making it well-suited for structured output tasks with strict metrics. However, broader application of GRPO can lead to entropy collapse, training instability, or diluted signal under multiple rewards. DAPO (Yu et al., 2025) addresses instability by maintaining output diversity, while GDPO (Liu et al., 2026) handles multi-reward settings by standardizing each reward independently before combination. We adopt GRPO and integrate DAPO and GDPO to mitigate these respective challenges.

2.4 Span Aggregation via Model Agreement

Individual models often disagree on span boundaries, so we aggregate predictions across models

to improve annotation reliability. Minimum Bayes Risk (MBR) decoding selects the output that maximizes agreement across a set of candidates, following a hypothesis that the output should be consistent with the others (Bickel and Doksum, 2015). It has shown to improve generation quality across NLP tasks. Heineman et al. (2024) adapted this intuition to multi-prompt decoding of a single model, leveraging varied inputs to encourage diverse prediction distributions. We adapt MBR decoding for clinical decision extraction, but extend prior work by aggregating outputs from multiple differently trained models and by performing refinement at the span-cluster level instead of choosing one complete prediction (Section 4.4).

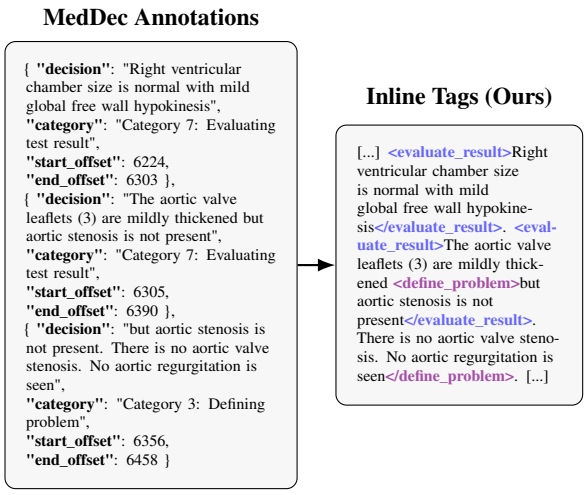

Figure 2: Converting original offsets (left) into inline tags that support containment and partial overlap (right).

3 Dataset

We use MedDec (Elgaar et al., 2024) as our dataset. For model development, we follow the original split of 350 training, 53 validation, and 48 test samples. The dataset consists of discharge summaries annotated with decision labels and span boundary offsets. We convert the original dataset into inline XML-like tags that support overlapping spans (Figure 2). This is necessary because 4.2% of tokens overlap. Therefore, this task encourages LLMs to reason through the input and apply inline-tag annotations where relevant.

Our work covers the 9 DICTUM classes defined in (Elgaar et al., 2026). The label distribution is highly imbalanced, with *Defining problem* accounting for 39% of training and validation spans, while *Deferment* accounts for under 0.2%. The dataset also has a demographic imbalance. For example, the Hispanic group contains just 1 sample each in validation and test, and Asian patients are entirely unrepresented in validation, with only 2 training samples (Appendix A). This makes the task challenging as it can strongly skew Worst Group F_1 scores and, in turn, affect Overall F_1. (Section 5.1).

During validation, a test run suggested a possible category misassignment in the ground-truth annotations. We locally reassigned the affected section to better reflect our model's true performance (Appendix B). Additionally, we used MIMIC-III discharge summaries outside the MedDec set for pseudo-label generation (Section 4.3).

4 Methodology

4.1 Dynamic Fine-tuning (DFT)

We train our first model variant (Model 1) using Dynamic Fine-Tuning (DFT) (Wu et al., 2026), which improves generalization by rescaling token gradients based on model confidence. Standard Supervised Fine-Tuning (SFT) applies a uniform cross-entropy loss, which often assigns disproportionately large gradients to tokens the model is uncertain about, leading to unstable optimization and overfitting. DFT addresses this by weighting each token's loss by its current generation probability. Formally, given an input sequence x and a target reference sequence $y^* = (y_1^*, \ldots, y_T^*)$ of length T, the DFT objective is defined as:

$$\mathcal{L}_{\text{DFT}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \text{sg}\big(\pi_\theta(y_t^* \mid y_{<t}^*, x)\big) \cdot \log \pi_\theta(y_t^* \mid y_{<t}^*, x), \quad (1)$$

where t is the current token index, π_θ is the model's policy parameterized by θ, $y_{<t}^*$ denotes the preceding target tokens $(y_1^*, \ldots, y_{t-1}^*)$, and $\text{sg}(\cdot)$ is the stop-gradient operator. This promotes balanced updates and encourages the model to learn robust extraction patterns rather than memorizing surface-level data.

4.2 Reinforcement Learning with Verifiable Rewards

4.2.1 Training Objective

We train our second model variant (Model 2) by initializing the policy with Model 1 (DFT), providing a robust starting point that allows the policy

to focus directly on optimizing extraction performance. We use a significantly reduced LoRA rank, exploiting the fact that policy gradient updates derived from sparse rewards inherently occupy a low-rank subspace. We optimize the policy using the Group Relative Policy Optimization (GRPO) (Shao et al., 2024) framework, which estimates advantages across sampled groups without a separate value model. We implement two targeted enhancements: First, we adopt the DAPO (Yu et al., 2025) approach at the token level to mitigate entropy collapse. By applying an asymmetric clipping range ($\epsilon_{low} = 0.20, \epsilon_{high} = 0.28$), DAPO safely broadens the trust region for stable exploration. Second, we integrate GDPO (Liu et al., 2026) to handle multi-objective rewards, standardizing each reward component independently across the group before summation. This preserves fine-grained distinctions between candidate generations and ensures each objective contributes equitably to the final gradient update.

4.2.2 Reward Functions

We translate our evaluation metrics into a composite reward system to optimize extraction performance. Relying on a single metric risks producing either sparse training signals or reward exploitation. To address this, we design a multi-objective reward that balances strict boundary evaluation, flexible word-level credit, and a hard fidelity constraint. Each component is bounded within $[0.0, 1.0]$, and the overall reward for each generation is defined as their combination:

1. **Fidelity Reward**: We design this component to strictly prevent hallucination. We strip all generated tags and compare the remaining text to the original input. Using a character-level similarity ratio, we assign a full reward of 1.0 for a near-perfect match where the ratio ≥ 0.99, and a reward of 0.0 for a ratio below 0.90, with linear scaling in between. This safeguards that the model behaves purely as a sequence tagger.

2. **Token F_1 Reward**: We use this component to provide flexible, word-level credit. We extract the predicted text spans, split them into individual words, and measure their overlap with the ground-truth words for each category. This allows the model to receive positive feedback for identifying relevant medical concepts even when exact span boundaries are slightly misaligned.

3. **Span F_1 Reward**: We treat this component as the primary objective and strictest metric. We evaluate whether the model precisely identifies the entire text span of a clinical decision, granting a reward only when the extracted string perfectly matches the ground truth within its category. By combining this strict target with the more forgiving Token F_1 reward, we guide the model toward predicting exact span boundaries.

4.3 Pseudo-label Generation

We train our third model (Model 3) by augmenting the training set with pseudo-labels generated from unlabeled clinical text. We identify semantically relevant unlabeled instances from 59,201 MIMIC-III discharge summaries outside of MedDec. We embed both the unlabeled pool and the complete MedDec set using Qwen3-Embedding-4B (Zhang et al., 2025). Then, we compute cosine similarity between each of the 451 MedDec documents and the unlabeled pool to retrieve the top 5 most similar candidates. We assign each unlabeled document only to its highest-scoring match, yielding a refined pool of 2,255 candidate samples. We then run inference on this pool using Qwen3.5-4B DFT (Model 1), using entropy as a proxy for prediction confidence based on the observed Pearson correlation of -0.4816 with Base Score on the validation set. A higher entropy threshold includes more examples but noisier labels. We experiment with two cutoffs at the 10th percentile (P_{10}) and 15th percentile (P_{15}), adding 221 and 331 pseudo-labeled samples, respectively. We combine each set with the original training set and fine-tune Model 3 using DFT.

4.4 Span-Cluster Consensus

We aggregate predictions by category and document in four stages. First, we greedily match spans above a minimum pairwise IoU threshold and compute pairwise soft Span F_1 agreement between models, then normalize these scores so that model weights sum to one. Second, we cluster all spans using a minimum cluster IoU threshold and score each cluster by the summed weights of its contributing models, removing clusters below the minimum cluster support. Third, we select final boundaries via a weighted vote over start–end offset pairs. Fourth, we merge same-category dupli-

cates above a minimum duplication IoU threshold, keeping the longer span. Figure 3 illustrates each stage.

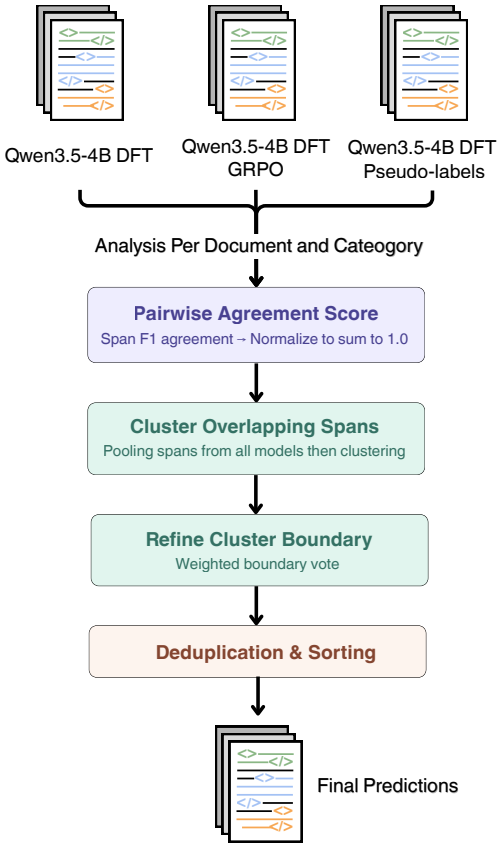

Figure 3: Overview of the Span-Cluster Consensus, comprising pairwise model agreement, cluster weighted boundary refinement, and span deduplication.

To find the best ensemble configuration, we performed a systematic search on the validation set over four fusion hyperparameters: minimum pairwise IoU, minimum cluster IoU, minimum cluster support, and minimum duplication IoU, each within a continuous range of [0.0, 1.0]. We used the Tree-structured Parzen Estimator (TPE) algorithm (Watanabe, 2023) with a warm-start trial seeded from our heuristic baseline to accelerate convergence. The search directly maximized the Overall F_1 score computed by the official evaluation script, ensuring that the selected configuration optimizes for both accuracy and demographic robustness jointly.

5 Experimental Setup

5.1 Evaluation Metrics

We evaluate with the official script and report Token F_1, Span F_1, Base Score, Worst Group F_1, and Overall F_1. Token F_1 measures word-level overlap on non-overlapping tokens, while Span F_1 requires exact span matches after word-boundary normalization. Both metrics are macro-averaged across the nine DICTUM labels per document and then across our internal validation or test set. Base Score is the mean of Token F_1 and Span F_1:

$$\text{Base Score} = \frac{\text{Span } F_1 + \text{Token } F_1}{2} \quad (2)$$

Worst Group F_1 is the lowest Base Score across sex, ethnicity, and language groups:

$$\text{Worst Group } F_1 = \min_{i \in \{1,\ldots,8\}} (\text{Base Score}_i) \quad (3)$$

and Overall F_1 is the average between Base Score and Worst Group F_1:

$$\text{Overall } F_1 = \frac{(\text{Base Score} + \text{Worst Group } F_1)}{2} \quad (4)$$

5.2 Baselines and Proposed Models

5.2.1 Encoder Baselines with BIO Tagging

As baselines, we evaluate a range of encoder-based models trained with BIO tagging for span detection, including models with stronger biomedical vocabulary coverage such as BioMedBERT (Gu et al., 2020), and models designed for longer clinical documents such as Clinical Longformer (Li et al., 2022). We also include ModernBERT (Warner et al., 2025), which incorporates architectural improvements such as RoPE positional encoding and an extended context window of 8,192 tokens. We additionally retrain RoBERTa and ELECTRA, the best-performing models from Elgaar et al. (2024), under our experimental configuration.

5.2.2 Generative Extraction Baseline

We evaluate two generative baselines. First, we fine-tune Qwen3.5-4B (Qwen Team, 2026) using standard SFT with autoregressive cross-entropy loss on the same training set and prompt as our proposed models (Appendix C). Comparing this baseline against Model 1 directly isolates the benefit of DFT over standard SFT. Second, we prompt GPT-5 (Singh et al., 2025) in a zero-shot setting using the same prompt, establishing an upper reference for generative extraction without task-specific training.

5.2.3 Proposed Models

We use Qwen3.5-4B (Qwen Team, 2026) as our base model for its reasoning capability and extended context window, which processes full discharge summaries without truncation. We train three variants based on strategies in Section 4:

1. **Model 1 (DFT Model):** We fine-tuned Qwen3.5-4B with the DFT objective, establishing it as a primary extraction model.

2. **Model 2 (DFT + RLVR Model):** We initialized from Model 1 and applied RLVR with DAPO and GDPO to further improve performance. Three reward combinations were tested: Span and Token F_1 (Model 2), Span F_1 only, and Token F_1 only. The first combination was also evaluated with metrics other than F_1, specifically $F_{0.5}$ and precision.

3. **Model 3 (DFT with Pseudo-label Model):** We fine-tuned Qwen3.5-4B with DFT on the original training set augmented with pseudo-labels generated by Model 1. We experimented with two entropy cutoffs including P_{10} (221 samples) and P_{15} (331 samples) to evaluate the tradeoff between pseudo-label quantity and quality.

Finally, predictions from all three models are aggregated using Span-Cluster Consensus (Section 4.4) and compared against two simpler baselines: majority voting (median start/end offsets, majority-voted category) and minimum-entropy selection (the single per-document prediction with lowest generation entropy).

5.3 Implementation Details

We implemented our pipeline using the Unsloth library (Daniel Han and team, 2023) and the Transformers Reinforcement Learning (TRL) framework (von Werra et al., 2020) across all training stages. All models use Qwen3.5-4B (Qwen Team, 2026) as the base model and AdamW (Loshchilov and Hutter, 2017) as the optimizer. Due to the differing computational demands, the DFT stages were trained on a single NVIDIA A100 (80GB) GPU, whereas RLVR was trained across four NVIDIA A100 (40GB) GPUs using Distributed Data Parallel (DDP) (Li et al., 2020). All models start with DFT via LoRA, where Model 1 and 2 were trained on a tag-converted training set and Model 3 with the addition of pseudo-labels. This stage used Rank-Stabilized LoRA (rsLoRA) (Kalajdzievski, 2023) with a rank of 256, $\alpha = 32$, 2 epochs, a learning rate of 2×10^{-4}, a batch size of 8, and a cosine learning rate scheduler. Model 2 further applies Reinforcement Learning from Verifiable Rewards (RLVR) trained for 1 epoch using the DAPO loss with standard LoRA, a lower learning rate of 5×10^{-5}, a larger batch size of 32, rollout size of 64, a constant scheduler, and a reduced LoRA rank and α of 2 (Schulman and Lab, 2025). For efficient inference, we utilized the vLLM engine (Kwon et al., 2023) with a temperature of 0. The ensemble hyperparameter search used the Optuna library with 5,000 trials using the default TPE sampler (Akiba et al., 2019).

6 Results

We evaluated various setups of model development across the Span F_1, Token F_1, and the composite scores. The following section reports the key results and their implications with respect to our internal validation set, unless specified otherwise. Moreover, our interpretation of the results focuses primarily on Base Score performance, since the Overall F_1 may be subject to the imbalanced split caused by the Worst Group F_1 as discussed in Section 3. Accordingly, we treat the Overall F_1 and Worst Group F_1 as supplementary context.

6.1 Encoder and Generative Baselines

All fine-tuned LLMs outperform both encoder and generative baselines (Table 1). Among encoder models, Clinical Longformer achieved the highest performance, with a Span F_1 of 0.4327, Token F_1 of 0.6297, and a Base Score of 0.5312. GPT-5, by contrast, yielded a competitive Span F_1 of 0.5179 but fell sharply on a Token F_1 of 0.2230 and a Base Score of 0.3210. Our LoRA SFT model outperforms both baselines across all three metrics, improving over Clinical Longformer by 1.96% in Token F_1 and 5.58% in Base Score, while also exceeding GPT-5 by 0.68% in Span F_1 and 26.60% in Base Score.

6.2 Fine-tuning Approaches

LoRA was the more efficient adapter, outperforming DoRA at the same rank and fine-tuning settings in both DFT (Base Score 0.5934 vs. 0.5826) and SFT (0.5870 vs. 0.5817) (Table 1). DFT also performed better than SFT in both LoRA (0.5934 vs. 0.5870) and DoRA (0.5826 vs. 0.5817). Although LoRA DFT had a lower Worst Group F_1 than LoRA SFT, the Worst Group F_1 for both methods

Approach	Model & Configuration	Span F_1	Token F_1	Base Score	Worst Group F_1	Overall F_1
Baselines	RoBERTa (BIO Tag)	0.4171	0.6233	0.5202	0.4416	0.4809
	ELECTRA (BIO Tag)	0.4061	0.5890	0.4976	0.4105	0.4540
	BioMedBERT (BIO Tag)	0.4126	0.5967	0.5047	0.4063	0.4555
	Clinical Longformer (BIO Tag)	0.4327	0.6297	0.5312	0.4609	0.4960
	ModernBERT (BIO Tag)	0.3282	0.5070	0.4176	0.3787	0.3981
Zero-shot	GPT-5-2025-08-07 (Medium Reasoning)	0.5179	0.2230	0.3210	0.3025	0.3118
Fine-tuning	Qwen3.5-4B (DoRA SFT, $r = 256$)	0.5250	0.6385	0.5817	0.4994	0.5406
	Qwen3.5-4B (DoRA DFT, $r = 256$)	0.5177	0.6475	0.5826	0.5151	0.5488
	Qwen3.5-4B (LoRA SFT, $r = 256$)	0.5247	0.6493	0.5870	<u>0.5440</u>	<u>0.5655</u>
	Model 1: Qwen3.5-4B (LoRA DFT, $r = 256$)	0.5299	0.6570	<u>0.5934</u>	0.5225	0.5580
Model Size	Qwen3.5-9B (LoRA DFT, $r = 256$)	0.5179	0.6582	0.5881	0.5044	0.5462
Rank Ablation	Qwen3.5-4B (LoRA DFT, $r = 128$)	0.5152	0.6464	0.5808	0.4716	0.5262
	Qwen3.5-4B (LoRA DFT, $r = 512$)	0.5273	0.6504	0.5889	0.4695	0.5292
RLVR Reward	Qwen3.5-4B (GRPO, Token F_1)	0.5258	<u>0.6577</u>	0.5917	0.5241	0.5579
	Qwen3.5-4B (GRPO, Span F_1)	0.5298	0.6573	0.5935	0.5261	0.5598
	Qwen3.5-4B (GRPO, Span and Token Precision)	0.5296	0.6523	0.5909	0.5249	0.5579
	Qwen3.5-4B (GRPO, Span and Token $F_{0.5}$)	0.5289	0.6560	0.5925	0.5309	0.5617
	Model 2: Qwen3.5-4B (GRPO, Span and Token F_1)	0.5273	0.6565	0.5919	0.5320	0.5619
Pseudo-labels	Qwen3.5-4B (Train set + P_{15})	0.5283	0.6460	0.5871	0.4783	0.5327
	Model 3: Qwen3.5-4B (Train set + P_{10})	**0.5307**	0.6481	0.5894	0.5022	0.5458
Ensemble	Majority vote	0.5298	0.6521	0.5910	0.5210	0.5560
	Minimum entropy	**0.5307**	0.6481	0.5894	0.5022	0.5458
	Final Model: Span-Cluster Consensus	<u>0.5300</u>	**0.6611**	**0.5955**	**0.5518**	**0.5737**

Table 1: Performance comparison of baselines, proposed configurations (**Model 1, 2, and 3**), ablations, and final ensemble on the internal validation set. **Bold** and <u>underline</u> indicate best and second-best results.

was the Hispanic subgroup, which contained only one sample and was therefore a less stable evaluation point. In all other subgroup analyses, LoRA DFT performed better than LoRA SFT. Therefore, we chose LoRA DFT as Model 1.

6.3 Adapter Rank and Model Size

Qwen3.5-4B was the optimal model size, outperforming Qwen3.5-9B (Base Score 0.5934 vs. 0.5881) (Table 1). At a fixed alpha of 32, an adapter rank of 256 achieved the highest Base Score (0.5934), compared with rank 128 (0.5808) and rank 512 (0.5889).

6.4 Reinforcement Learning Configuration

The Token F_1-only reward variant achieved the highest Token F_1 score (0.6577), while the Span F_1-only variant achieved the best Span F_1 score (0.5298) (Table 1). Notably, the Span F_1-only variant also maintained a comparatively high Token F_1 score (0.6573), resulting in the highest Base Score among the variants at 0.5935. In addition to F_1, we explored other evaluation metrics, including $F_{0.5}$, which weights precision more heavily than recall, and precision, which measures the fraction of predicted positive instances that are truly positive. Among these metrics, $F_{0.5}$ appears to perform best, followed by F_1 and then precision, with Base Scores of 0.5925, 0.5919, and 0.5909, respectively.

6.5 Pseudo-labels Thresholds

When defined as the entropy percentile threshold, P_{10} outperformed P_{15} across all metrics (Table 1). Model 3 (P_{10}) achieved the highest Span F_1 of 0.5307 and a Token F_1 of 0.6481, while also misclassifying fewer spans as non-decision, particularly in minority classes such as *Gathering additional information*, *Treatment goal*, and *Deferment*.

6.6 Ensemble Models

The ensemble outperformed the individual components and all other variants on all but one metric, achieving the best Token F_1 (0.6611), Base Score (0.5955), Worst Group F_1 (0.5518), and Overall F_1 (0.5737) (Table 1). Its Span F_1 of 0.5300 was the second highest, narrowly behind Model 3 (0.5307).

Rank	Team	Span F_1	Token F_1	Base	Worst F_1	Overall
1	billbaumgartner	0.5419	0.6667	0.6043	0.5886	0.5965
2	**LAMAR (Ours)**	0.5257	0.6750	0.6003	0.5881	0.5942
3	Otter	0.5181	0.6666	0.5924	0.5695	0.5809
4	viahes	0.5237	0.6541	0.5889	0.5723	0.5806
5	ahmed_ayman	0.4900	0.6796	0.5848	0.5601	0.5724
-	Baseline	0.4363	0.6238	0.5301	0.4922	0.5111

Table 2: Official Leaderboard of MedExACT Shared Task (Elgaar et al., 2026). Our proposed system (LAMAR) achieves highly competitive performance, securing 2nd place overall.

6.7 Test Set Performance

Our system ranked second on the test leaderboard (Table 2). Compared with the first-place system, we achieved an Overall F_1 that was 0.23% lower, but a 0.83% higher Token F_1. We also consistently outperformed the baseline across all scores with a 7.02% improvement on Base Score. As shown in Table 3, our method performs consistently well across all demographic subgroups with the Worst Group F_1 at 0.5881 in the Non-English subgroup.

Category	Subcategory	Span F_1	Token F_1	Base Score
Gender	Male	0.5117	0.6678	0.5897
	Female	0.5495	0.6850	0.6173
Ethnicity	White	0.5216	0.6726	0.5971
	African American	0.4962	0.6926	0.5944
	Hispanic	0.5761	0.6794	0.6277
	Asian	0.5319	0.7090	0.6205
	Other	0.5538	0.6651	0.6094
Language	English	0.5303	0.6853	0.6078
	Non-English	0.5156	0.6605	0.5881

Table 3: Test set performance of our proposed system across demographic subgroups.

7 Qualitative and Error Analysis

Our system missed approximately 40% of ground truth spans overall, and around 30% of non-decision spans were misclassified as *Defining problem*. Minority classes were the most difficult to predict. The most extreme case was Model 2 (DFT + GRPO), which correctly identified only 2 out of 41 *Gathering additional information* spans (4.88%). This likely reflects the influence of class imbalance across data splits (Appendix A) on the model's behavior. The most common decision-to-decision misclassification was *Evaluating test results* being predicted as *Defining problem*, occurring at a rate of 6.8 – 9.1%. We hypothesize that this is due to the fact that evaluating a test result often requires inferring the current state of a condition, thereby blurring the boundary between test-result evaluation and defining clinical problems. The qualitative example from the validation set is illustrated in Figure 4.

8 Discussion

The results collectively demonstrate that encoder baselines perform well on Token F_1 by design, while zero-shot GPT-5 highlights that LLMs can effectively extract span-level information but still struggle in token-level settings. Inline-tag fine-tuning bridges this gap, as even LoRA SFT outperforms the baselines across metrics. Among the fine-tuning strategies, Qwen3.5-4B LoRA with DFT rank 256 consistently emerged as the strongest single-model configuration. The reinforcement learning results behaved as theoretically expected, with each reward type optimizing its corresponding metric, reinforcing the importance of aligning reward signals with evaluation objectives. The pseudo-label experiments further revealed sensitivity to label noise: a tighter entropy threshold (P_{10}) produced cleaner pseudo-labels and meaningfully improved minority-class recall. Finally, the ensemble's strong overall performance, driven by complementary prediction patterns between models trained with DFT, DFT + GRPO, and DFT with pseudo-labels demonstrates that ensembling effectively compensates for individual model weaknesses. Test-set results further show that the system generalizes consistently across demographic subgroups, with the Worst Group F_1 not arising from disproportionately small subgroups.

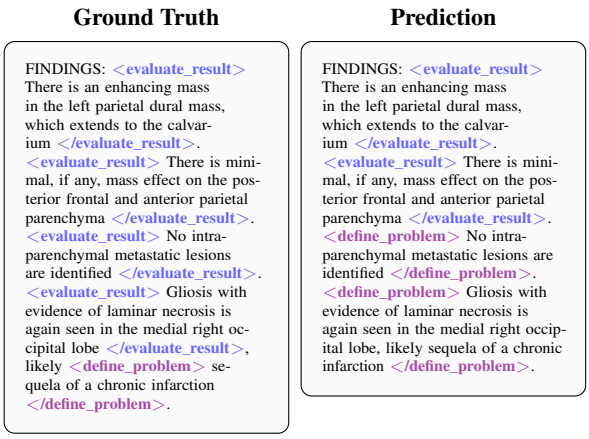

Figure 4: Most common decision-to-decision misclassification: *Evaluating test results* being predicted as *Defining problem*.

9 Conclusion

Our work demonstrates that the agreement-driven ensemble consistently outperforms both encoder-based and zero-shot generative baselines across all metrics. This suggests that combining predictions from models with different output distributions can better capture complementary representations. We further show that performance can be improved through fine-tuning with DFT, reinforcement learning with verifiable rewards, and pseudo-label augmentation, highlighting the value of combining complementary modeling strategies with targeted training enhancements.

Limitations

Our study has several limitations spanning methodology, data, and experimental scope. First, full-document generation with inline tags is more computationally expensive at inference time than classifier-based BIO tagging, which may limit practicality in clinical settings. Although we use a fidelity reward to reduce hallucinations, this mechanism has not been tested across diverse inputs, and its robustness remains uncertain. Second, our experiments are limited to Qwen3.5-4B and a single dataset, MedDec (451 documents), leaving generalizability an open question. Third, we observed overlapping spans within the same category, including cases with slightly different offsets and cases with identical boundaries. Cleaning these annotations led to worse performance than retaining the original data, likely because overlapping spans are also present in the validation set. This issue warrants further investigation. Fourth, pseudo-label thresholding remains a challenge. Our results show slightly negative correlations, suggesting the current strategy does not reliably distinguish useful pseudo-labels from noisy ones. Because entropy captures only model confidence, it may be insufficient as the sole filtering criterion. Finally, time constraints limited our exploration of reinforcement learning policies, and the current setup may not reflect the optimal policy or ensemble configuration. Although Model 2 shows notable potential for reward-based improvement, this study prioritized ensemble methods, and a deeper investigation of reward configuration is left for future work.

Acknowledgments

We thank Pattaramanee Arsomngern for her thoughtful feedback on this manuscript, and to Kunat Pipatanakul and Chompakorn Chaksangchaichot for their discussions on model development.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Peter J Bickel and Kjell A Doksum. 2015. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

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, volume 1 (long and short papers)*, pages 4171–4186.

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.

Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Mehrnaz Sadrolashrafi, Mitra Mohtarami, Adrian Wong, Hadi Amiri, and Leo A. Celi. 2026. Overview of medical decision extraction, analysis, and classification task (medexact) 2026. In *The 25th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, San Diego, California, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

David Heineman, Yao Dou, and Wei Xu. 2024. Improving minimum bayes risk decoding with multi-prompt. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22525–22545.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, and 1 others. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

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. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*.

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*.

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*.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and 1 others. 2020. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704*.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. 2026. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *Preprint*, arXiv:2601.05242.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

John Schulman and Thinking Machines Lab. 2025. Lora without regret. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/lora/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. Gpt-ner: Named entity recognition via large language models. In *Findings of the association for computational linguistics: NAACL 2025*, pages 4257–4275.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Shuhei Watanabe. 2023. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2026. On the generalization of SFT: A reinforcement learning perspective with reward rectification. In *The Fourteenth International Conference on Learning Representations*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Exploratory Data Analysis

We examined the demographic breakdown of the train, validation, and test splits. Table 4 highlights the severe underrepresentation of some groups, particularly across ethnicities, which can substantially affect the Worst Group F_1 score. Table 5 provides additional context on the prevalence of each decision category in the dataset, offering insight into observed model behavior.

B Internal Validation Set

We reassigned all spans from a section shown in Figure 5 from *Treatment goal* to *Defining problem* in the original validation set for internal validation. The original span boundaries were preserved, and only the category labels were changed.

C Decision Extraction Prompt

We designed the inline tag extraction prompt shown in Figure 6. This prompt was used for supervised fine-tuning, RLVR fine-tuning, and GPT-5 zero-shot evaluation.

D Encoder Model Training Configurations

For the encoder baselines, we adopted MedDec's sliding-window training and non-overlapping window inference strategy while modifying several training hyperparameters. The window size for each model was set to its maximum supported input length: 512 for RoBERTa, ELECTRA, and BioMedBERT; 4096 for Clinical Longformer; and 8192 for ModernBERT. With these settings, RoBERTa, ELECTRA, and BioMedBERT were trained on a single NVIDIA A100 80GB GPU, whereas Clinical Longformer and ModernBERT required greater computational resources and were trained using 4 NVIDIA A100 40GB GPUs. All models were trained for 250 epochs with a learning rate of 2×10^{-5}, an effective batch size of 32, and a cosine learning rate scheduler.

Category	Subcategory	Train	Validation	Test
Gender	Male	204	27	28
	Female	146	26	20
Ethnicity	White	244	43	35
	African American	31	7	4
	Hispanic	21	1	1
	Asian	10	0	2
	Native Hawaiian	1	0	0
	Other	43	2	6
Language	English	197	35	28
	Non-English	153	18	20

Table 4: Dataset distribution by category and subcategory across train, validation, and test sets.

Class	Train	%Train	Validation	%Validation
Contact related	2293	5.28	289	4.12
Gathering additional information	376	0.87	41	0.58
Defining problem	17015	39.19	2766	39.39
Treatment goal	107	0.25	46	0.66
Drug related	10930	25.17	1989	28.33
Therapeutic procedure related	5311	12.23	825	11.75
Evaluating test results	5806	13.37	828	11.79
Deferment	84	0.19	10	0.14
Advice and precaution	1496	3.45	228	3.25

Table 5: Number of spans per class in the train and validation sets.

Excerpt of the Internal Validation Set

[...] **Review of systems:** (+) Per HPI (-) Denies fever, chills, night sweats, recent weight loss or gain. Denies headache, sinus tenderness, rhinorrhea or congestion. Denies cough, shortness of breath, or wheezing. Denies chest pain, chest pressure, palpitations, or weakness. Denies nausea, vomiting, diarrhea, constipation, abdominal pain, or changes in bowel habits. Denies dysuria, frequency, or urgency. Denies arthralgias or myalgias. Denies rashes or skin changes.

Past Medical History: EtOH Abuse Cirrhosis Hepatitis C: No prior treatment Diabetes Mellitus 2 - 20 + years Tobacco Use Depression Hypertension GERD Pancreatitis Diverticulitis Hemorrhoids Atypical chest pain

Social History: - Tobacco: 1 ppd x 20+ years - Alcohol: 6-12 beers daily - Illicits: None

Family History: No history of bleeding disorders or abdominal bleeding. Both parents still living. **Physical Exam: Vitals:** T: 97 BP: 127/54 P: 112 R: 18 18 O2: 96/RA **General:** Alert, oriented, no acute distress **HEENT:** Sclera anicteric, MMM, oropharynx clear **Neck:** supple, JVP not elevated, no LAD **Lungs:** Clear to auscultation bilaterally, no wheezes, rales, ronchi **CV:** Regular rate and rhythm, normal S1 + S2, no murmurs, rubs, gallops **Abdomen:** soft, non-tender, non-distended, bowel sounds present, no rebound tenderness or guarding, no organomegaly **GU:** no foley **Ext:** warm, well perfused, 2+ pulses, no clubbing, cyanosis or edema [...]

Figure 5: Reassigned spans in the internal validation set. Only spans in this section originally categorized as *Treatment goal* were reassigned to *Defining problem*.

Decision Extraction Prompt

You are an expert specializes in extracting clinical decisions from a patient's discharge summary.

YOUR TASK

Given an input discharge summary, return the EXACT SAME text, but with specific phrases wrapped in inline tags to mark clinical decisions.

IMPORTANT:

- Do NOT add, remove, or rephrase any text outside the tags.
- Preserve all original punctuation, line breaks, and spacing.
- EVERY opening tag MUST have a corresponding closing tag (e.g., <drug_decision>Aspirin 81 mg daily</drug_decision>).
- These tags CAN overlap or nest in one another, as long as they are VALID TAGS.

DECISION CATEGORIES & TAGS

Use the following tags exactly as defined:

1. <define_problem> : diagnostic conclusions, health state evaluations, etiological inference, or prognostic judgment.

2. <drug_decision> : decisions to start, stop, continue, withhold, or modify medications.

3. <evaluate_result> : interpretation of clinical findings or test results.

4. <contact_related> : admissions, discharges, follow-ups, or referrals to other hospitals.

5. <therapeutic_procedure> : decisions to perform, plan, or refrain from procedures.

6. <advice_and_precaution> : patient instructions, advice, or precautions.

7. <gather_info> : decisions to order tests and investigations or consult another colleague.

8. <treatment_goal> : therapeutic goals, aims, or treatment objectives.

9. <defer_decision> : delaying judgment or action for now.

ANNOTATION RULES

1. Boundary: Annotate spans that capture the full clinical decision. Prefer longer spans than short words.

2. Comprehensiveness: The output should be comprehensively annotated. Extract as many valid decisions as possible.

3. Exclusions: DO NOT annotate document headers or labels (e.g., "Admission Date:", "Discharge Date:", "Physical Exam:").

4. Overlapping Spans: Spans may overlap or belong to multiple categories. Wrap each span independently with all applicable tags.

Example (nested): <drug_decision>continue warfarin for <treatment_goal>stroke prevention</treatment_goal></drug_decision>

Example (partial overlap): <define_problem>The next previous examination suggested <evaluate_result>atelectasis</define_problem> - density in the left base cannot be evaluated</evaluate_result>

OUTPUT FORMAT

Return ONLY the fully annotated text. Ensure all tags are properly closed. Do not include any explanations.

INPUT TEXT

{discharge_summary}

Figure 6: Decision extraction prompt used in our system and for zero-shot evaluation of GPT-5.