Active-Learning–Guided Scoring Model for Rare-Entity Acquisition in Clinical NER

Anonymous ACL submission

Abstract

Adverse Drug Events (ADEs) are a major cause of preventable morbidity and mortality, underscoring their importance for clinicians and patient safety initiatives. Yet ADE mentions in clinical narratives are rare, context dependent, and frequently multi-token, which makes them difficult to capture using standard active learning (AL) heuristics. We introduce a meta-model-driven AL framework that predicts, for each unlabelled sentence, a proxy performance gain by combining uncertainty signals, embedding-level diversity, ontology alignment, and a multi-word ontology-match (MWOM) cue. A random forest surrogate ranks sentences by predicted utility. On n2c2 2018 Track 2 with BioClinicalBERT-CRF, the method annotates 1,250 of 3,949 pool sentences (32%), for a total of 5,256 labels versus 7,955, yet achieves entity-level micro F1 of 0.9781 (mean over five seeds; 95% CI [0.9774, 0.9788]) and macro F1 of 0.8069, exceeding the strongest non meta baseline (R Cos: micro 0.9601, macro 0.6599). Rare type gains are large: ADE 0.0352 to 0.3057 (8.68×), Reason 0.3712 to 0.5689, Duration 0.1276 to 0.7230. These results indicate that learned, feature rich acquisition better targets rare, safety critical entities while reducing annotation effort, offering a practical route to scalable pharmacovigilance in clinical NER.

1 Introduction

Named-entity recognition (NER) in electronic health records (EHRs) is essential for converting free text into structured clinical information such as medications, ADEs, and their attributes (Meystre et al., 2008). Although deep neural networks achieve high accuracy when trained on large annotated corpora (Collobert et al., 2011), clinical annotation remains costly because it must be performed by domain experts (Sarker and Gonzalez, 2015). AL seeks to mitigate this cost by selecting the most informative samples for labelling (Settles, 2008). However, widely used single-heuristic selection rules such as uncertainty sampling, query-by-committee, or diversity maximisation often underperform in clinical NER. The reason is not merely the complexity of clinical text but the difficulty of defining informativeness when the learner must decide simultaneously (i) whether an entity is present, (ii) where its span begins and ends, (iii) which of several fine-grained types to assign, and (iv) how that decision interacts with other entities in the same sentence. Token-level confidence scores capture only a fragment of this joint space, and current span-level uncertainty estimators are noisy, especially for rare, multi-word ADE mentions.

We therefore propose an AL strategy guided by a meta-model that predicts, for every unlabelled sentence, the expected improvement in downstream F1 if that sentence were annotated. The meta-model combines complementary signals, including token-entropy uncertainty, embedding-level semantic diversity, ontology alignment, and a multi-word ontology match (MWOM) cue for long and rare spans. Trained on historical or simulated acquisitions, it outputs a PPG that enables ranking by estimated utility rather than by any single heuristic. Under a fixed four-round budget, this approach attains strong overall accuracy with substantially fewer labels and markedly better coverage of rare entities.

Our contributions are as follows:

- We introduce a learned selector that predicts a per-sentence PPG using a meta-model trained on simulated acquisition deltas, replacing static heuristics with an adaptive utility estimate aligned with downstream F1.

- We unify three complementary signals by combining uncertainty, embedding-level diversity, and ontology cues within a single feature space, including MWOM that targets long

and rare spans and systematically surfaces ADE, Reason, and Duration mentions.

- We achieve efficiency and accuracy under a fixed four-round budget by annotating only 1,250 of 3,949 pool sentences (32%), for a total of 5,256 labels (34% fewer than 7,955), while reaching micro-F1 0.9781 (five-seed mean; 95% CI [0.9774, 0.9788]), outperforming the strongest non-meta baseline by +1.80 micro-F1.

- We improve rare-entity coverage, raising ADE F1 from 0.0352 to 0.3057 (8.68×), with parallel gains for Reason (+0.1977) and Duration (+0.5954); acquisition-behaviour analysis shows that queried batches are consistently enriched for these spans.

To our knowledge, this is among the first meta model driven AL frameworks for clinical NER that learns a PPG surrogate and integrates ontology aware MWOM with uncertainty and diversity cues within a single predictive model.

2 Related Work

2.1 AL Cuts Annotation Cost-But With Caveats

Early pool-based AL studies on generic sequence labelling showed that query-by-committee and uncertainty sampling can slash annotation budgets by 40-80% compared with random selection (Settles, 2012; Tomanek and Hahn, 2010). When first ported to clinical notes, the same heuristics delivered substantial savings (up to 66% fewer labelled sentences) but also exposed two limitations: strong class imbalance and domain-specific terminology made rare concepts hard to acquire (Chen et al., 2015). Subsequent work incorporated token-level cost models: Wei et al. (2019b) showed that a cost-aware sampler Cost-CAUSE improved the area-under-learning-curve by 5-6 pp while cutting annotation time by approximately 25%.

2.2 Rare but Safety-Critical Entities

Clinical stakeholders are more concerned about low-frequency ADEs. The results from the 2018 n2c2 ADE shared task showed that even the best current methods still missed many rare mentions of ADE (Henry et al., 2020b). Follow-up work (Henry et al., 2022) demonstrated that classical uncertainty sampling under-samples such uncommon ADEs, degrading recall once the prevalence falls below 3%. These findings motivate AL strategies that explicitly prioritise long-tail risk signals.

2.3 Transformer Era: Better Calibration, Same Static Heuristics

Domain-specific transformers such as Clinical-Longformer Gao et al. (2024) and Clinical ModernBERT Lee et al. (2025) produced markedly better probability calibration on long notes, yielding more reliable uncertainty scores in low-label regimes. Yet most AL pipelines still plugged these models into a static acquisition rule, typically entropy or margin sampling, leaving the weighting between uncertainty, diversity and ontology cues frozen for the entire project.

2.4 Static Multi-Cue Samplers

Several recent systems enriched entropy with fixed auxiliary scores. BEAL combined Bayesian uncertainty with confidence estimation for multi-label text classification (Wang and Liu, 2023); ACSESS linearly re-weighted 20 cues via LASSO on a dev set (Pecher et al., 2024); and ontology-guided heuristics injected concept density from clinical knowledge bases to accelerate early recall (Alluri et al., 2025). While these methods outperformed vanilla entropy in the first few AL rounds, performance often plateaued once the underlying label distribution drifted (Kohl et al., 2024).

2.5 Towards Dynamic Acquisition Strategies

A growing body of work began to adapt the selection criterion on-the-fly. Liu et al. (2025) proposed a switchable uncertainty–diversity strategy that reduced human edits per concept when stricter quality thresholds were applied. Luo et al. (2023) smoothed token utilities with dynamic reweighting, stabilising selection under severe label sparsity. Beyond clinical NLP, dynamic loss-based selection (Yoo and Kweon, 2019) and core-set selection (Sener and Savarese, 2018) demonstrated similar benefits in vision tasks, while Monte-Carlo-Dropout uncertainty (Gal and Ghahramani, 2016) offered a Bayesian alternative for query scoring. Nevertheless, existing dynamic schemes typically adjusted one cue at a time and did not jointly optimise heterogeneous signals.

Motivation for This Study The literature demonstrated (i) the clear promise of AL for cutting annotation cost in clinical NER and (ii) persistent

shortcomings when sampling relied on static or single-cue criteria, especially for rare ADEs. We therefore propose a multi-cue, meta-model that re-optimises cue weights at every AL round, learning a proxy for expected performance gain and sustaining improvements throughout the entire annotation trajectory.

3 Methodology

3.1 Model Architecture

We used a BERT–CRF architecture for clinical NER. The encoder was BioClinicalBERT (Alsentzer et al., 2019), which provided contextual token representations. A linear layer projected hidden states to label logits, and a conditional random field (CRF) (Lafferty et al., 2001) modeled label dependencies to enforce valid BIO tag transitions. WordPiece tokenisation was used; labels were assigned to the first subtoken and propagated to subsequent subtokens as I-tags, and special tokens were masked in the loss. The model was trained with sequence level negative log likelihood under the CRF. Training and optimisation details, including learning rate schedule, batch size, and early stopping, are given in Section 4.5.

3.2 AL Strategies

To ensure a fair comparison of all selection strategies, we followed a four-round AL schedule. Rounds 1–3 queried 250 sentences each (about 6% of the 3,949-sentence CLiPS pool), based on class-frequency estimates suggesting that, at the observed 2–3% ADE prevalence, a batch of this size would likely yield at least five new ADE spans. This quantity was sufficient to help recalibrate the model's handling of rare entities while keeping annotation demands manageable and feedback cycles responsive. In the final round, a larger batch of 500 sentences (about 13%) was selected to consolidate earlier gains and target remaining ADE-rich examples in a single sweep, while still preserving a reserve of unlabelled data. In total, we annotated 1,250 sentences (32% of the CLiPS pool), balancing reduced human effort with broad and diverse coverage. A detailed description of the dataset partitioning appears in Section 4.3.

Active-learning heuristics such as uncertainty and diversity sampling have repeatedly proved useful in clinical NER, though they have not been systematically evaluated on the n2c2 2018 ADE corpus (Henry et al., 2020a). To establish strong baselines, we re-implemented four prominent strategies under an identical setup. First, U-MCD (Gal and Ghahramani, 2016) estimated model uncertainty via stochastic inference: each sentence was passed through the model twenty times with dropout activated, token-level entropies were computed from the mean predictive distribution, and then averaged to a sequence-level score. Second, U-Ent (Settles, 2008) used the same acquisition function but from a single deterministic forward pass (no dropout). Third, D-Cos (Yoo and Kweon, 2019) greedily selected sentences whose [CLS] embeddings were maximally dissimilar (cosine) from those in the labelled pool, promoting semantic diversity. Fourth, R-Cos (Sener and Savarese, 2018) used a k-center greedy objective to select sentences that best covered the embedding distribution of the unlabelled pool. While all four offer well-founded heuristics, they are not explicitly tuned for infrequent, domain-specific entities such as ADEs. To address this limitation, we proposed a meta-model-driven strategy that combines uncertainty estimates, ontology-aware matching, and semantic diversity cues to guide annotation. Gold labels were programmatically retrieved from the fully annotated n2c2 2018 dataset to simulate human annotations while eliminating inter-annotator variability and ensuring controlled experimentation.

Span-count diagnostic. For acquisition-behaviour analysis we counted rare spans by scanning gold labels for B-tags of {ADE, Reason, Duration} in each selected sentence (continuations I-belong to the same span).

3.3 Meta-Feature Extraction and Predicted Proxy Gain

To improve the effectiveness of instance selection in our AL methodology, we introduced a meta-model-based strategy that predicts a PPG for each unlabelled sentence. The PPG serves as an estimate of how much the inclusion of a given instance, once annotated, is expected to boost model performance. Unlike conventional approaches that rely solely on uncertainty-based heuristics, our method incorporates a diverse set of domain-informed meta-features, capturing uncertainty, ontology alignment, semantic diversity, and concept density.

Uncertainty was quantified using token-level entropy derived from the CRF's marginal probability distributions. For each token t, the model outputs a predicted label distribution $\mathbf{p}_t =$

$[p_t, 1, \ldots, p_t, L]$, where L denotes the number of possible labels. The entropy at each token is then computed as

$$H_t = -\sum _l = 1^L p_t, l \log p_t, l, \quad (1)$$

and the overall sentence-level uncertainty is obtained by averaging across all tokens:

$$H_\text{seq} = \frac{1}{T} \sum _t = 1^T H_t, \quad (2)$$

where T is the number of tokens. This approximation sidestepped the intractability of computing exact CRF sequence entropy while still offering an effective measure of uncertainty in AL contexts.

To complement uncertainty, we incorporated ontology-guided binary features that indicated whether the predicted sequence contained ADE or Reason mentions and whether these aligned with known ontology entries. We defined four binary indicators: ADE presence ($f1$), Reason presence ($f2$), ADE ontology match ($f3$), and Reason ontology match ($f4$). Each feature was set to 1 if the condition was satisfied and 0 otherwise:

$$f_i = \begin{cases} 1, & \text{if the feature is present in the instance,} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

These features provided a lightweight, interpretable signal that helped prioritise sentences likely to contain relevant clinical concepts.

Semantic diversity was captured through a global novelty score based on the [CLS] embedding of the sentence. We computed the Euclidean distance between this embedding and the centroid of all [CLS] embeddings in the current labelled pool:

$$d_\text{cls} = \|\mathbf{e}_\text{CLS} - \mathbf{c}_\text{train}\|, \quad (4)$$

where $\mathbf{e}$_CLS represents the current sentence embedding and $\mathbf{c}$_train is the mean embedding over the labelled data. While local alternatives such as nearest-neighbour distances could capture more fine-grained novelty, we opted for the centroid-based approach due to its stability and its utility in early rounds when the labelled pool was sparse.

Finally, we included domain-specific density cues by counting the number of predicted ADE-related and Reason-related terms in each sentence. These scalar features, denoted as $f5$ and $f6$, helped prioritise sentences that were conceptually rich and thus more likely to yield informative annotations for low-frequency classes.

To balance the influence of different feature types, we applied standard score normalisation to each real-valued meta-feature:

$$\hat{f}_j = \frac{f_j - \mu_j}{\sigma_j}, \quad (5)$$

where μ_j and σ_j were the mean and standard deviation of feature f_j within the current labelled set. This yielded a standardised feature vector $\mathbf{z} = [\hat{f}_1, \ldots, \hat{f}_8]$ for each unlabelled instance.

We evaluated three variants of the PPG function. The first, M-ADE, was a heuristic that combines entropy with ADE/Reason binary signals:

$$\text{PPG_M–ADE} = H_\text{seq} + 0.5 \sum_{i=1}^{4} \mathbf{1}(f_i). \quad (6)$$

The second, M-FocusADE, downweighted entropy and focused exclusively on presence indicators for ADE and Reason:

$$\text{PPG}_{\text{M–FocusADE}} = -H_{\text{seq}} + 0.5 \left(\mathbb{1}(f_1) + \mathbb{1}(f_2) \right). \quad (7)$$

Both heuristics were interpretable and computationally efficient. The third variant, M-Optimal, used a random forest regressor $\mathcal{R}$ to map the full feature vector to a predicted gain:

$$\widehat{\text{PPG}} = \mathcal{R}(\mathbf{z}). \quad (8)$$

In addition to the core meta-features, M-Optimal uniquely incorporated a MWOM indicator. This feature used a context-aware n-gram strategy (window size ± 3) to match phrases against curated ADE and Reason lexicons, thereby enhancing the detection of long-span, domain-specific entities. The regressor was retrained after each AL round, enabling adaptive refinement as new annotations were incorporated.

3.4 Random-Forest Regression Model

To implement M-Optimal, we benchmarked several regressors and selected the random forest based on its empirical performance. In preliminary experiments, the random forest consistently produced stronger F1 improvements for underrepresented entities such as ADE and Duration compared to alternatives like XGBRegressor (Appendix B). We employed the scikit-learn implementation with 300 trees, unrestricted depth, and a minimum leaf size of one, using the squared-error loss as the splitting criterion. A fixed random seed (42) ensured reproducibility across runs.

Training involved simulating a full active-learning cycle using the unlabelled pool. Each instance was assigned a gold-standard PPG (see Appendix A for details), and its corresponding meta-features were normalised to produce the training matrix.

The gold-standard PPG label quantified the expected utility of annotating each instance during simulation. Specifically, when an instance x_i (or its batch) was selected and added to the labelled set, we measured the resulting change in validation micro-F1 (ΔF1) relative to the previous round. This ΔF1 served as the gold PPG target for x_i, reflecting its empirical contribution to model improvement under controlled conditions. The random forest regressor $\mathcal{R}$ was then trained to approximate this mapping from the normalised meta-feature vector $\mathbf{z}_i$ to the scalar $\widehat{\mathrm{PPG}} = \mathcal{R}(\mathbf{z}_i)$, enabling data-driven ranking of unlabelled sentences by predicted performance gain.

We then split this matrix into 80% training and 20% validation data using `train_test_split`. This held-out validation set remained constant across rounds, providing a stable benchmark to monitor generalisation. While finer-grained evaluations could have been conducted per round, we prioritised efficiency and clarity by focusing on end-point assessments.

3.5 Integration into the AL Loop

In each round of the AL cycle, the trained random forest predicted a PPG score for every unlabelled instance in the pool. Sentences were then ranked by their predicted scores, and the top k instances (250 in the first three rounds and 500 in the final) were selected for annotation. Once annotated, these instances were appended to the training set, meta-features were recomputed, and the regressor was retrained to reflect the expanded labelled pool.

To enable incremental learning across active cycles, the model underwent parameter updates rather than full retraining from scratch. Newly annotated instances from each cycle were programmatically appended to the training set via PyTorch's `ConcatDataset` mechanism, progressively expanding the knowledge base while preserving prior annotations.

Figure 1 depicts the full pipeline. Following initial fine-tuning of the BERT-CRF model, each AL round proceeded through the following steps: (1) sample unlabelled data, (2) extract standardised meta-features encompassing uncertainty, diversity, and ontology alignment, (3) predict proxy gains using the current regressor, (4) rank and select the top sentences for annotation, and (5) retrain the NER model on the updated labelled dataset. This design facilitated adaptive learning while maintaining procedural consistency.

The experiments in Section 4 evaluated three main questions derived from this methodology: (i) whether the M-Optimal strategy better identifies rare, high-value sentences than standard baselines; (ii) whether our fixed four-round budget delivers meaningful gains in annotation efficiency; and (iii) whether the BERT-CRF model remains stable under incremental updates. These questions are addressed using performance metrics such as micro/macro F1 and annotation efficiency, with special focus on low-frequency clinical entity types.

4 Experiments

4.1 Dataset Overview and Preprocessing

This study utilised the n2c2 2018 Track 2 challenge dataset, a benchmark corpus for extracting medication-related information from EMRs (Henry et al., 2020b). The dataset consists of annotated clinical notes containing entities such as drugs, ADEs, reasons for medication use, and medication attributes (e.g., dosage, strength, frequency, and route).

The dataset was originally provided in unstructured text format with inline entity annotations. To enhance compatibility with transformer-based architectures, it was converted into a structured token-level format.

4.2 Named Entity Annotation Schema

The BIO tagging format was adopted for sequence labelling tasks, where each token in a sentence was assigned a label indicating its role in an entity span (Ramshaw and Marcus, 1995; Tjong Kim Sang and De Meulder, 2003). This format includes three types of labels: B (Begin), which marks the first token of an entity; I (Inside), which is used for subsequent tokens within the same entity; and O (Outside), which designates tokens that do not belong to any named entity.

4.3 Data Splitting and Preprocessing

The dataset used in this study consisted of 299 training files and 201 test files, with the test set retained in its original format for evaluation. To

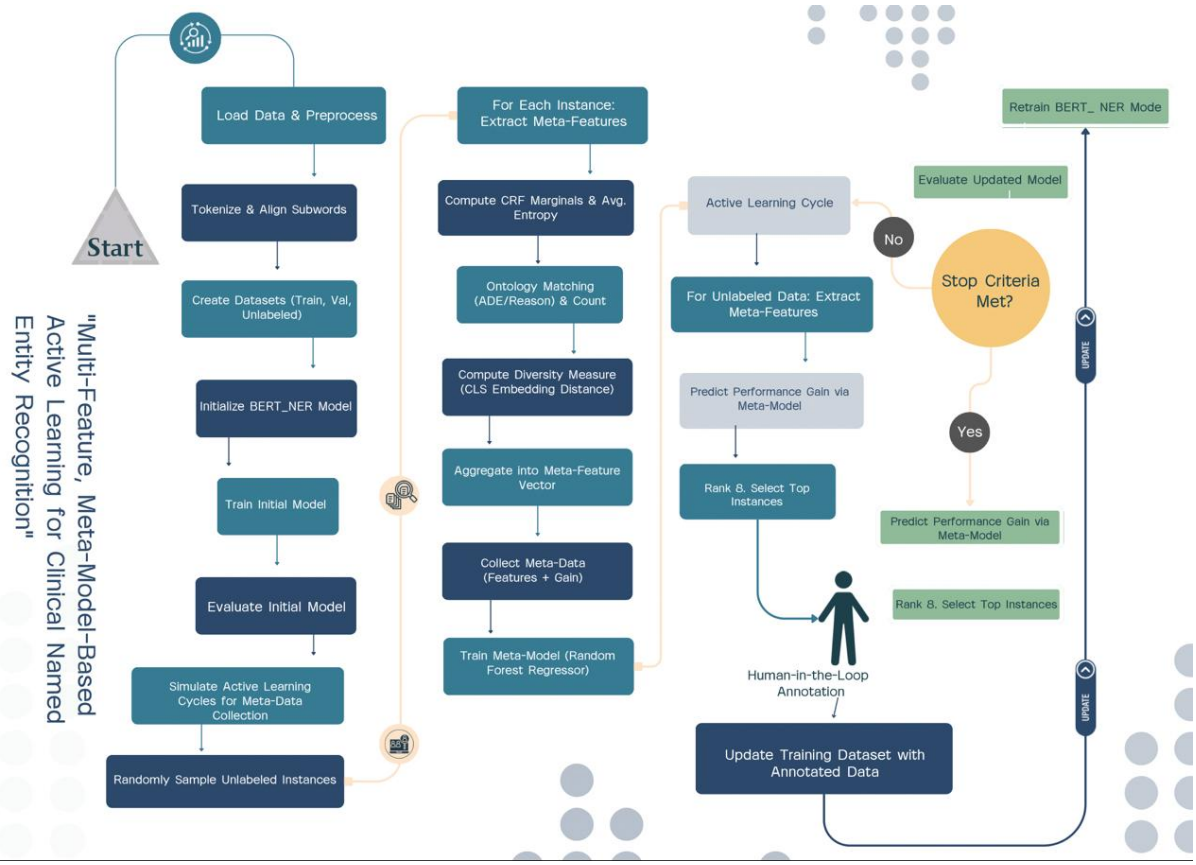

Figure 1: Overview of the meta-model-driven AL pipeline. The figure outlines the process of meta-feature extraction, proxy gain computation, and integration into the AL loop. **Human annotation steps are simulated by programmatically retrieving ground truth labels from the held-out CLiPS subset, ensuring consistency and reproducibility across experiments.**

ensure a diverse and representative training distribution, we applied a distance-aware stratification strategy using K-Means clustering with $k = 10$. This approach grouped sentences based on their semantic similarity, ensuring a balanced mix of both representative and rare cases. By leveraging this clustering mechanism, we enhanced the generalisation capability of the model by exposing it to a varied set of linguistic patterns.

The initial labelled set comprised 4006 sentences, while the CLiPS (unlabelled pool) contained 3949 sentences. For clarity, while our dataset contains 808,163 token-level entries, these are organised into 15,284 unique sentences. Specifically, the dataset was partitioned into 4006 sentences for training, 1249 for validation, 3949 for the CLiPS pool, and 6080 for testing. This structured organisation facilitated controlled experimentation and consistent evaluation across different AL cycles.

4.4 Evaluation Metrics

Performance evaluation was conducted using micro and macro F1-scores, precision, recall, annotation efficiency, and confusion matrix analysis. The micro F1-score measured the overall proportion of correctly classified tokens across all classes, whereas the macro F1-score computed the average F1-score per class, providing insight into the model's ability to generalise across different entity types. Precision and recall metrics further assessed the trade-off between false positives and false negatives. Annotation efficiency was introduced to quantify the improvement in F1-score per annotated instance, allowing for an evaluation of how effectively the selected samples contributed to model performance. A confusion matrix was also generated to analyse common misclassification patterns, offering valuable insights into systematic errors and potential areas for refinement.

4.5 Training and Optimisation

We fine-tuned BioClinicalBERT with a CRF decoder using AdamW (learning rate 2×10^{-5}, weight decay 10^{-2}), batch size 16, dropout 0.1, and maximum sequence length 105. We apply cosine annealing with warm restarts (T_0 = 5, T_{mult} = 1, $\eta_{\min} = 10^{-6}$) and gradient clipping at 1.0. Training uses early stopping on validation F1 with a patience of five epochs. The initial fit runs for up to 15 epochs, and each AL cycle also trains for up to 15 epochs under the same criterion. To mitigate catastrophic forgetting during iterative learning, we integrated Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), a regularisation technique that penalised substantial deviations from previously learned parameters.

Multi-seed protocol. To assess robustness, we evaluated M-Optimal under multiple random seeds ({42, 123, 456, 789, 101}) using the AL schedule defined in Section 3.2 (four cycles; total 1,250 additional annotations). Multi-seed evaluation was applied only to M-Optimal to conserve compute, as it is the primary method of interest.

5 Results

Overall and per-entity trends. Under a fixed budget of 1,250 sentences (32% of 3,949 CLiPS), all models were fine-tuned from the same 4,006 annotated sentences and evaluated on the 201 original test files. The strongest non-meta baseline was R-Cos micro-F1 0.9601. Both M-ADE 0.9702 and M-Optimal surpassed it; M-Optimal achieved micro-F1 0.9781 (five seeds: 0.9780, 0.9780, 0.9788, 0.9787, 0.9770; SD 0.0006; 95% CI [0.9774, 0.9788]) and macro-F1 0.8069 vs. 0.6599 for R-Cos, reflecting improved coverage of both frequent and rare tags. At the entity level, the static baseline performed adequately on frequent classes (e.g., Dosage 0.8204, Drug 0.8809) but poorly on low-prevalence types; non-meta AL nudged these up max ADE 0.0352. In contrast, M-Optimal substantially improved rare types: ADE 0.3057, Reason 0.5689, Duration 0.7230, consistent with combining uncertainty, semantic diversity, and MWOM cues (Table 2).

See Appendix D for a comparison of AL strategies across entity types (Figure 3).

Acquisition behaviour. We counted gold spans of ADE, Reason, and Duration in each queried batch using B-tag occurrences. Table 1 reports per-round counts for seed 42.[1] Aggregated per 250 sentences, M-Optimal selects approximately 246 ADE, 880 Reason, and 196 Duration spans, compared with Uncertainty (209, 800, 109) and Diversity (61, 323, 56), aligning with the observed gains on these types. Residual confusions are concentrated between ADE and O, and between Reason and Drug; the full confusion matrix and discussion are provided in Appendix C.

Seed	Round	Method	k	ADE	Reason	Duration
42	1	M-Optimal	250	77	250	55
42	1	Diversity	250	19	96	12
42	1	Uncertainty	250	52	246	25
42	2	M-Optimal	250	90	231	50
42	2	Diversity	250	24	87	15
42	2	Uncertainty	250	88	207	34
42	3	M-Optimal	250	48	227	54
42	3	Diversity	250	10	87	17
42	3	Uncertainty	250	45	200	21
42	4	M-Optimal	500	61 (30.5)	343 (171.5)	74 (37)
42	4	Diversity	500	16 (8.0)	105 (52.5)	24 (12)
42	4	Uncertainty	500	48 (24.0)	294 (147)	58 (29)

Table 1: Rare-span coverage in the selected batches (seed 42). Round 4 used k=500; per-250 equivalents are shown in parentheses.

Annotation efficiency. Despite annotating only 1,250 sentences from the CLiPS pool (versus 3,949 under full pool annotation), M-Optimal improved ADE F1 from 0.0352 (R-Cos) to 0.3057, an 8.68× increase, and raised macro F1 by 0.147 (0.8069 versus 0.6599). These gains indicate that the meta model effectively directs limited expert effort toward high yield sentences.

Comparison to prior work. Table 3 situates our results on the n2c2 2018 ADE dataset: M-Optimal attains the highest micro-F1 while using fewer labels (5,256 vs. 7,955).

6 Discussion

The findings indicate that a learned acquisition surrogate aligned with downstream utility can systematically address rare-type blind spots in clinical NER. Rather than querying sentences that are merely uncertain or globally diverse, the meta-model prioritises instances whose feature profiles (uncertainty, semantic distance, ontology alignment, and MWOM) historically correlate with gains in held-out F1. This explains both the rare-type improvements and the macro-level gains at

[1] Diagnostic only; not used for significance testing. In Round 4, k=500; per-250 equivalents are shown in parentheses.

Tag	B-Init	U-MCD	D-Cos	R-Cos	U-Ent	M-ADE	M-Optimal	M-FocusADE
ADE	0.0024	0.0224	0.0235	0.0352	0.0024	0.2306	***0.3057***	0.1894
Dosage	0.8204	0.8569	0.8709	0.8692	0.8779	0.9250	***0.9268***	0.9214
Drug	0.8809	0.9282	0.9302	0.9202	0.9282	***0.9502***	0.9502	0.9471
Duration	0.0276	0.0276	0.0376	0.1276	0.0276	0.7124	***0.7230***	0.6801
Form	0.5918	0.8324	0.8424	0.8354	0.8424	***0.9077***	0.9051	0.9039
Frequency	0.7243	0.8493	0.8594	0.8403	0.8523	0.8914	0.8832	***0.8915***
O	0.9746	0.9826	0.9803	0.9836	0.9820	***0.9850***	0.9850	0.9849
Reason	0.1855	0.3745	0.3768	0.3712	0.3794	0.5527	***0.5689***	0.5583
Route	0.6375	0.7823	0.7856	0.7892	0.7819	0.8582	0.8614	***0.8640***
Strength	0.8079	0.9112	0.9178	0.9124	0.9138	***0.9609***	0.9599	0.9576
Micro F1	0.8563	0.9585	0.9592	0.9601	0.9585	0.9702	***0.9781***	0.9699
Macro F1	0.5653	0.6586	0.6588	0.6599	0.6566	0.7974	***0.8069***	0.7898

Table 2: F1 scores for baseline, non-meta, and meta-model strategies (0 denotes non-entity tokens). Bold-italic marks the best value in each row. Per-tag scores are from a representative seed (42); the M-Optimal Micro F1 (0.9781) is the mean over five seeds.

Model	Year	Type	Micro–F1	Annotated Sentences
Co-occurrence rule-based (Mahendran and McInnes, 2021)	2021	Rule-based	0.86	7,955
Segment-CNN (Mahendran and McInnes, 2021)	2021	CNN-based	0.78	7,955
BioBERT fine-tuned (Mahendran and McInnes, 2021)	2021	BERT-based	0.94	7,955
ClinicalBERT fine-tuned (Wei et al., 2019a)	2019	BERT-based	0.94	7,955
M-Optimal (Ours)	**2025**	**Meta-model (AL)**	**0.9781**	**5,256**

Table 3: Comparison of micro–F1 performance and annotation sizes for models evaluated on the 2018 n2c2 ADE dataset. Our method uses significantly fewer annotated sentences while outperforming all baselines on micro–F1.

a fixed budget: acquisition-behaviour diagnostics show batches enriched with ADE, Reason, and Duration spans (Table 1), which mirrors the per-entity gains (Table 2). The micro/macro trends are stable across five seeds under identical training and evaluation settings, and the efficiency claim is supported by higher yield per annotated sentence relative to non-meta baselines and prior systems trained with more labels (Table 3). Residual errors concentrate around ADE vs. O boundary cases and Reason vs. Drug overlaps; the full confusion matrix and analysis are provided in Appendix C.

See Section 8 for scope and threats; immediate extensions include matched multi-seed baselines with paired tests, structured meta-feature ablations, span-aware objectives to reduce boundary and overlap errors, and faster pool scoring via cached embeddings or approximate nearest-neighbor search.

7 Conclusion

We introduced a meta-model-driven AL framework for clinical NER that predicted a PPG for each unlabelled sentence by integrating uncertainty, semantic diversity, ontology alignment, and an MWOM cue. The selector was trained on gold-standard PPG labels, measured as validation ΔF1 during simulated acquisitions, enabling it to prioritise sentences with the highest expected contribution to model improvement.

Across five random seeds, M-Optimal achieved a mean micro F1 of 0.9781 (standard deviation 0.0006; 95% confidence interval [0.9774, 0.9788]) while annotating 1,250 of 3,949 pool sentences (32%). Counting the full training plus CLiPS set, the final model used 5,256 labelled sentences, which is 34% fewer overall than 7,955. Macro F1 reached 0.8069, exceeding the best non meta baseline (0.6599), and rare type gains were large for ADE, Reason, and Duration.

Future work will include cross dataset validation to assess generalisability, structured ablations to isolate meta feature contributions, span aware decoders to address boundary errors, and efficiency improvements to reduce selection latency in large pools.

8 Limitations

Single-dataset scope. Results are currently reported only on the n2c2 2018 ADE corpus. Evaluations on additional clinical datasets and under domain shift are required to establish external validity.

Seed comparability across methods. Multi-seed evaluation was conducted for M-Optimal to as-

sess robustness; baseline strategies were executed under identical splits but not replicated across seeds due to computational constraints. A more rigorous comparison would include matched seeds for all baselines and paired statistical tests.

Ablation of meta-features. The observed gains are attributed to uncertainty, semantic diversity, and MWOM cues, supported by acquisition diagnostics. However, a complete ablation isolating each component's marginal contribution remains pending. Future work will incorporate structured ablations and permutation-importance analyses.

Simulated annotation regime. Annotations were programmatically retrieved from gold labels to minimise variability. While standard in AL simulations, this design does not capture human-in-the-loop factors such as correction latency, adjudication effort, or annotator fatigue. Prospective studies measuring annotation time and quality would strengthen empirical claims about efficiency.

Boundary and overlap errors. Confusions persist between ADE and non-entities, and between Reason and Drug, reflecting span-boundary and semantic-overlap challenges. Promising solutions include span-level decoders, contrastive objectives to separate overlapping types, and ontology-guided hard-negative mining.

Computational overhead. Meta-feature extraction entails multiple embedding passes over large pools. Prototype-based indexing, cached embeddings, or approximate nearest-neighbour search could alleviate latency without degrading selection quality.

Potential risks. Although the study uses de-identified clinical data, deployment of medical NER systems must consider risks such as mislabelling or omission of adverse events, which could lead to misinformation or biased downstream analyses. Human oversight and external validation are essential prior to clinical application.

References

Shailesh Alluri, Trupti Chitale, and Kristina Vopat. 2025. An active-learning pipeline to identify candidate terms for a clinical decision support ontology. In *Studies in Health Technology and Informatics*, volume 316, pages 1338–1342. IOS Press. PMID: 40321267.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Yebo Chen, Thomas A. Lasko, Qiang Wei, Joshua C. Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of Biomedical Informatics*, 58:11–18.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 48:1050–1059.

Qian Gao, Mingyuan Li, and Byron Wallace. 2024. Clinicallongformer: Efficient full-context self-attention for long clinical notes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '24)*, pages 3124–3138, Mexico City, Mexico. Association for Computational Linguistics.

Nicholas Henry, Zulfat Miftakhutdinov, and Isaac Hochberg. 2022. Uncertainty sampling under-samples rare adverse drug events in electronic health records. *Journal of the American Medical Informatics Association*, 29(10):1748–1757.

S. Henry, K. Buchan, M. Filannino, A. Stubbs, Ö. Uzuner, and E. Soysal. 2020a. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 1–10. Association for Computational Linguistics.

Stephanie Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Özlem Uzuner. 2020b. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Philipp Kohl, Yoka Krämer, Claudia Fohry, and Bodo Kraft. 2024. Comparative performance analysis of active learning strategies for the entity recognition task. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR)*, pages 480–488.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence

data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289. Morgan Kaufmann.

James J. Lee, Wonjin Cho, and Joonhwan Lee. 2025. Clinical modernbert: An efficient long-context encoder pre-trained with umls descriptors. *arXiv preprint*.

... Liu and 1 others. 2025. Utilizing active learning strategies in machine-assisted annotation for clinical named entity recognition. *JAMIA*, 31(11):2632–2644.

H. Luo, W. Tan, N. Dang Nguyen, and L. Du. 2023. Re-weighting tokens: A simple and effective active learning strategy for named entity recognition. *arXiv preprint arXiv:2311.00906*.

Darshini Mahendran and Bridget T. McInnes. 2021. Extracting adverse drug events from clinical notes. *AMIA Joint Summits on Translational Science Proceedings*, 2021:420–429.

Stephane M. Meystre, Guergana K. Savova, Kipper-Schuler K. C., and John F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics*, 47(1):128–144.

Branislav Pecher, Ivan Srba, Maria Bielikova, and Joaquin Vanschoren. 2024. ACSESS: Automatic combination of sample-selection strategies for few-shot learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP '24)*, pages 1234–1246, Bangkok, Thailand. Association for Computational Linguistics.

L. A. Ramshaw and M. P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 82–94.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196–207.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*.

Burr Settles. 2008. Active learning literature survey. In *University of Wisconsin-Madison Department of Computer Sciences Technical Report*.

Burr Settles. 2012. *Active Learning*, volume 6 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Katrin Tomanek and Udo Hahn. 2010. A semi-supervised scenario for learning base noun phrase recognition. *ACM Transactions on Speech and Language Processing*, 7(2):3:1–3:23.

Yu Wang and Yang Liu. 2023. BEAL: Bayesian expected-confidence active learning for multi-label text classification. In *Proceedings of the 2023 Conference on Findings of ACL*, pages 2450–2463, Toronto, Canada. Association for Computational Linguistics.

Qi Wei, Zhu Ji, Yuan Si, Jiang Du, Jun Wang, Fikret Tiryaki, Stephen Wu, Cui Tao, Kirk Roberts, and Hua Xu. 2019a. Relation extraction from clinical narratives using pre-trained language models. In *AMIA Annual Symposium Proceedings*, volume 2019, pages 1236–1245.

Qiang Wei, Yukun Chen, Mandana Salimi, Joshua C. Denny, Qiaozhu Mei, Thomas A. Lasko, Qingxia Chen, Stephen Wu, Amy Franklin, Trevor Cohen, and Hua Xu. 2019b. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322.

Donggeun Yoo and In So Kweon. 2019. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 93–102.

A Meta-Feature Representation

This appendix provides a detailed breakdown of the meta-feature representation used in AL. Meta-features capture various aspects of each instance, such as uncertainty (entropy), the presence of adverse drug events (ADEs), semantic diversity, and ontology-based matches.

Gold PPG Assignment. During the simulated AL cycle, each acquired batch of sentences was evaluated on a held-out validation set. We computed the change in validation micro-F1 between successive rounds and assigned this value as the gold-standard PPG to all instances within the batch. This batch-level approximation provides a practical surrogate for per-instance utility while avoiding exhaustive retraining on every candidate. The resulting pairs of meta-feature vectors and gold PPG labels formed the training data for the random-forest regressor described in Section 3.4.

The following tables illustrate example meta-feature values used during acquisition. ✓ indicates feature presence and ✗ indicates absence. The proxy gain is a learned score predicted by the random-forest regressor; instances with higher values are prioritised for annotation.

Sentence	Avg Entropy	B-ADE	I-Reason	ADE Onto	Reason Onto	Proxy Gain (Example)
Patient experienced severe nausea after taking aspirin.	0.78	✓	✗	✓	✗	0.65
Aspirin caused an allergic reaction in a 45-year-old patient.	0.85	✓	✓	✓	✓	0.85
The drug did not cause any side effects.	0.40	✗	✗	✗	✗	0.15
Paracetamol is used for pain relief.	0.35	✗	✗	✗	✗	0.12
Ibuprofen caused stomach ulcers in some patients.	0.92	✓	✓	✓	✓	1.1
Antibiotics led to severe diarrhea in elderly patients.	0.88	✓	✓	✓	✓	0.88

Table 4: Example meta-features used for AL. ✓ indicates presence and ✗ indicates absence. The proxy gain is a learned score: a random-forest regressor estimates it from the input meta-features and higher values are prioritised for annotation.

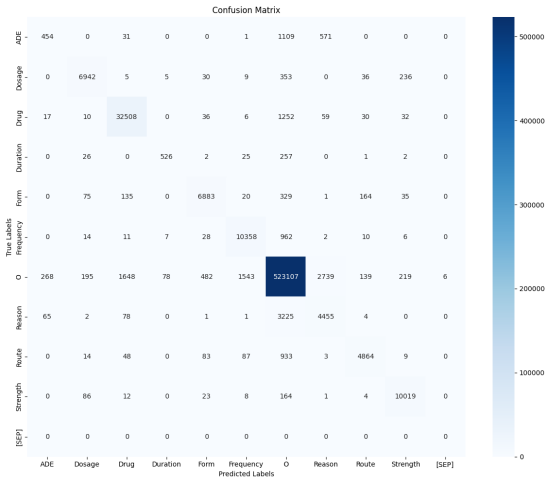

Figure 2: Confusion matrix after the final AL cycle.

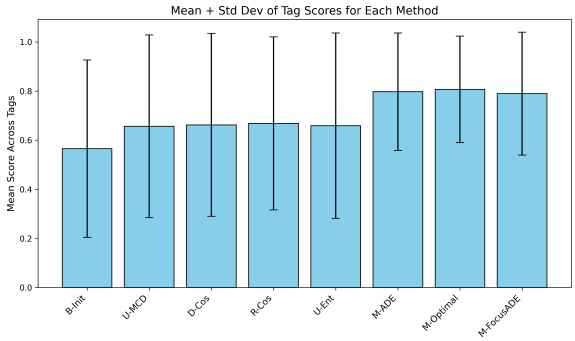

Figure 3: Performance comparison of AL strategies across different entity types. Error bars denote standard deviation.

B Comparison of Regression Models

To support the selection of Random Forest as the meta-regressor, Table 5 compares its F1 performance (as part of the M-Optimal strategy) with that of XGBRegressor, evaluated after AL on the test set. Random Forest outperformed XGBoost particularly on rare and domain-specific entities such as ADE and Duration, supporting its use for prioritising informative examples in clinical NER.

C Confusion Matrix and Analysis

Residual errors are dominated by ADE↔O and Reason↔Drug, indicating boundary ambiguity and semantic overlap. These patterns are consistent with the main text results.

D Additional Results: Performance by Entity

Figure 3 summarises the per-entity outcomes. Meta-model-driven methods achieve higher average F1 scores overall; M-Optimal stands out, likely due to the inclusion of MWOM and meta-features that capture uncertainty, diversity, and domain-specific cues. Larger standard deviations across methods reflect the difficulty of recognising rare entities like ADE and Duration.

Entity Type	M-Optimal (Random Forest)	XGBRegressor
ADE	**0.3057**	0.1820
Dosage	**0.9268**	0.9240
Drug	**0.9502**	0.9474
Duration	**0.7230**	0.6844
Form	**0.9051**	0.8934
Frequency	0.8832	**0.9012**
Reason	**0.5689**	0.5519
Route	**0.8614**	0.8553
Strength	**0.9599**	0.9545
Micro F1	**0.9781**	0.9702
Macro F1	**0.8069**	0.7163

Table 5: Comparison of F1-scores between Random Forest (M-Optimal) and XGBRegressor across entity types. Random Forest showed clear advantages, particularly for rare classes (e.g., ADE, Duration).