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, yet ADE mentions in clinical narratives are rare, context-dependent, and often span multiple tokens, which limits the effectiveness of standard active learning (AL) heuristics. We propose a meta-model–driven AL framework that ranks unlabelled sentences by a predicted proxy performance gain (PPG), estimated from uncertainty signals, embedding-level diversity, ontology alignment, and a multi-word ontology match (MWOM) cue. A random forest regressor is used as a surrogate utility model to combine these features.

Evaluated on the n2c2 2018 Track 2 dataset, the proposed method selects 1,250 of 3,949 candidate sentences (32%), resulting in 5,256 annotated sentences in total, compared with 7,955 sentences under full-pool annotation. Across five random seeds, the method achieves a mean Micro F1 of 0.9475 (95% CI [0.9469, 0.9480]), while attaining a Macro F1 of 0.7871 in a single-seed analysis (seed 42), outperforming the strongest non–meta-model baseline (R-Cos: Micro 0.9311, Macro 0.6334). Substantial gains are observed for rare entity types in the same run, with ADE improving from 0.0352 to 0.3057, Reason from 0.3712 to 0.5689, and Duration from 0.1276 to 0.7230. These results demonstrate that learned, feature-rich acquisition strategies can more effectively prioritise rare, safety-critical entities while substantially reducing annotation requirements.

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 proxy performance gain (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. (2019) 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: Long-Context Encoders, Still Fixed Query Rules

Recent domain-specific encoders extend context length for clinical text, reducing truncation and enabling richer representations of long notes (e.g., Clinical-Longformer/Clinical-BigBird (Li et al., 2022) and Clinical ModernBERT (Lee et al., 2025)). However, in much of the NLP AL literature, stronger encoders are still typically paired with a fixed acquisition rule across rounds (often entropy, margin, or MC-dropout variants), rather than adapting how uncertainty, diversity, and domain cues are traded off as the labelled pool grows (Zhang et al., 2022; Schröder et al., 2022; Vacareanu et al., 2024; Shelmanov et al., 2021). This is not merely a modelling choice: practical analyses with transformers emphasise that sophisticated pool-scoring objectives can be computationally prohibitive, which partly explains why uncertainty-based heuristics remain common baselines in real pipelines (Schröder et al., 2022).

2.4 Static Multi-Cue Samplers

A second line of work augments uncertainty sampling with additional, pre-specified signals. For example, BEAL proposes an acquisition criterion for deep multi-label text classification that combines Bayesian uncertainty estimation with an expected-confidence objective (Wang et al., 2024). More broadly, ACSESS demonstrates that combining multiple selection strategies can be beneficial, but also reports strong dependence on dataset and regime (e.g., number of shots), suggesting that a single fixed recipe may not be uniformly optimal (Pecher et al., 2024). Taken together, these results motivate moving beyond hand-tuned, fixed-weight combinations, especially in structured prediction settings where the usefulness of uncertainty and representativeness can vary substantially over AL rounds (Zhang et al., 2022; Schröder et al., 2022; Vacareanu et al., 2024). However, because these combinations are fixed once defined, they cannot adapt to changes in class prevalence or model behaviour across AL rounds.

2.5 Towards Dynamic Acquisition Strategies

A growing body of work explicitly adapts the selection criterion across rounds. In clinical NER, dynamic strategies that switch or blend diversity- and uncertainty-based sampling have been evaluated under simulated machine-assisted annotation, with attention to annotation costs and target effectiveness (Liu and Wong, 2024). For token-level structured prediction, dynamic re-weighting of token utilities has also been proposed to stabilise acquisition under sparsity and imbalance (Luo et al., 2023). Despite this progress, many dynamic schemes adjust only a narrow subset of cues (e.g., uncertainty vs. diversity) and do not jointly integrate structured domain knowledge. Critically, existing dynamic approaches typically adapt only a narrow subset of signals (e.g., uncertainty versus diversity) and do not learn a unified utility model that integrates uncertainty, representation-level information, and structured clinical knowledge.

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-estimates the relative utility of heterogeneous cues at each 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 candidate pool, denoted CLiPS), 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.

Upfront annotation. Here, an instance refers to a sentence. In a real setting, the initial training set (4,006 sentences) and validation set (1,249 sentences) must be manually annotated before AL begins; in our simulation we reveal the corresponding gold labels.

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. B-Init (initial model baseline). B-Init denotes the performance of the base NER model trained only on the initial labelled set (4,006 sentences), before any active-learning acquisition rounds. It serves as a reference point for the absolute gains achieved by subsequent acquisition strategies under the same architecture, training schedule, and evaluation protocol.

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 (co-

sine) 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 these four heuristics are well motivated, they are not explicitly designed to prioritise low-prevalence, domain-specific entities such as ADEs. To separate the effect of acquiring annotation-dense sentences from the effect of meta-model learning, we additionally include a density-only baseline that ranks unlabelled sentences using concept-density cues (i.e., the counts of predicted ADE/Reason terms, f_5+f_6) and selects the top-k sentences per round under the same four-round budget. Building on this, we propose a meta-model-driven strategy that combines uncertainty estimates, ontology-aware matching, and semantic diversity cues to guide acquisition. In all active-learning simulations, annotations are revealed by programmatically retrieving the corresponding gold labels from the fully annotated n2c2 2018 dataset, eliminating inter-annotator variability and enabling controlled comparisons.

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

$$H_t = -\sum_{l=1}^{L} p_{t,l} \log p_{t,l}, \tag{1}$$

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

$$H_{\text{seq}} = \frac{1}{T} \sum_{t=1}^{T} H_t, \tag{2}$$

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 (f_1), Reason presence (f_2), ADE ontology match (f_3), and Reason ontology match (f_4). Each feature was set to 1 if the condition was satisfied and 0 otherwise:

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

These features provided a lightweight, interpretable signal that helped prioritise sentences likely to contain relevant clinical concepts. We computed a novelty score based on the [CLS] embedding of the sentence by measuring its Euclidean distance from the centroid of all [CLS] embeddings in the current labelled pool:

$$d_{\text{cls}} = \left\| \mathbf{e}_{\text{CLS}} - \mathbf{c}_{\text{train}} \right\|, \tag{4}$$

where $\mathbf{e}_{\text{CLS}}$ represents the current sentence embedding and $\mathbf{c}_{\text{train}}$ is the mean embedding over the labelled data. This score captures how atypical a candidate is relative to the labelled pool (i.e., global novelty), rather than enforcing pairwise diversity within the selected batch. 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 f_5 and f_6, 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:

$$\hat{f}_j = \frac{f_j - \mu_j}{\sigma_j}, \tag{5}$$

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:

$$\mathrm{PPG}_{\text{M–ADE}} = H_{\text{seq}} + 0.5 \sum_{i=1}^{4} \mathbb{1}(f_i). \tag{6}$$

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

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

Here, entropy is negated to prioritise low-entropy (high-confidence) sentences that already exhibit ADE or Reason signals, reflecting an exploitation-oriented strategy that reinforces rare-entity decision boundaries rather than exploring highly uncertain or noisy instances.

Both heuristics were interpretable and computationally efficient but relied on fixed weights that do not adapt across AL rounds. The third variant, M-Optimal, used a random forest regressor $\mathcal{R}$ to map the full feature vector to a predicted gain:

$$\widehat{\mathrm{PPG}} = \mathcal{R}(\mathbf{z}). \tag{8}$$

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 D). 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 a query batch of k sentences was selected in an AL round and added to the labelled set, we measured the resulting change in validation micro-F1 (ΔF1) relative to the previous round. We used this round-level ΔF1 as a batch-level proxy-gain target and assigned it to each sentence in the selected query batch. The random forest regressor $\mathcal{R}$ was then trained to map 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.

Following initial fine-tuning of the BERT-CRF model, each AL round proceeds through: (1) scoring the unlabelled pool, (2) extracting standardised meta-features (uncertainty, diversity, ontology cues), (3) predicting proxy gains with the current regressor, (4) ranking and selecting the top-k instances for annotation, and (5) updating the NER model on the expanded labelled set. A full pipeline schematic is provided in Appendix B (Figure 1).

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

Why n2c2 2018 Track 2. We choose n2c2 2018 Track 2 because it is a standard clinical NER benchmark that jointly targets ADEs, medication attributes, and indication-like Reasons in EHR narratives, matching our long-tail, multi-span extraction setting. Its established train–test split and evaluation protocol enable controlled, reproducible comparison of acquisition strategies under identical budgets. We leave cross-corpus validation to future work (Section 8).

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 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 entity-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

We evaluate NER using span-based precision, recall, and F1, where a prediction is counted as correct only if it forms an entity span with a valid begin and end boundary under the BIO scheme and is assigned the correct entity type. Following the n2c2 Track 2 evaluation protocol, we report lenient span-level scores in the main paper: a predicted span is considered a match if it overlaps the gold span and the entity type agrees. Micro F1 aggregates true positives, false positives, and false negatives over all entity types before computing F1, reflecting overall extraction performance. Macro F1 is computed as an unweighted mean of per-type F1 across the nine entity types (ADE, Dosage, Drug, Duration, Form, Frequency, Reason, Route, Strength), providing a balanced view of performance on both frequent and low-prevalence entities. We additionally analyse annotation efficiency and error patterns

via span-level confusion analysis.

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 acquisition budget of 1,250 sentences (32% of the 3,949-sentence CLiPS pool), all models were fine-tuned from the same 4,006 annotated sentences and evaluated on the 201 original test files. Among non-meta strategies, R-Cos was strongest (Micro F1 0.9311; Macro F1 0.6334). To test whether gains arise mainly from selecting annotation-dense sentences (via f_5/f_6), we add a Density-only baseline (f_5+f_6); it remains well below M-Optimal overall (Micro F1 0.9225 vs 0.9484) and on rare types (ADE 0.0992 vs 0.3057; Reason 0.2779 vs 0.5689; Duration 0.0514 vs 0.7230; Table 1). Both meta strategies improved upon this baseline, with M-ADE reaching Micro F1 0.9413 and Macro F1 0.7766. M-Optimal was strongest overall in the representative run (Micro F1 0.9484; Macro F1 0.7871; Table 1). Across five random seeds, M-Optimal achieved a mean Micro F1 of 0.9475 (0.9470, 0.9470, 0.9478, 0.9478, 0.9478; SD 0.0004; 95% CI [0.9469, 0.9480]), indicating stable gains under different random initialisations. At the entity level, the static baseline performed adequately on frequent classes (e.g., Dosage 0.8204, Drug 0.8809) but remained weak on low-prevalence types; non-meta AL improved these only modestly (e.g., ADE up to 0.0352). In contrast, M-Optimal substantially improved rare types, notably ADE 0.3057, Reason 0.5689, and Duration 0.7230, consistent with combining uncertainty, semantic diversity, and MWOM cues (Table 1).

Acquisition behaviour. M-Optimal tends to acquire batches enriched with rare spans (ADE/Reason/Duration), consistent with the rare-type gains reported in Table 1. Detailed per-round coverage statistics (seed 42) are provided in Appendix B.1.

Annotation cost (true effort). Because sentence counts can mask substantial variation in annotation effort, we additionally analysed the true gold annotation cost of the M-Optimal acquisitions in terms of entity spans and entity tokens. Full cost statistics under the fixed acquisition schedule are reported in Appendix C.

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. At the aggregate level, Macro F1 increased by 0.154 (0.7871 versus 0.6334), reflecting a systematic shift toward better coverage of low-prevalence entity types under the same acquisition budget.

Comparison to prior work. Table 2 contextualises our end-point performance on the same benchmark under lenient evaluation. Although these prompt-based MRC systems are not active-learning baselines (they are trained on the full labelled training set), M-Optimal attains a comparable lenient Micro F1 (0.9475; five-seed mean) while using fewer annotated sentences overall (5,256 versus 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

Tag	B-Init	U-MCD	D-Cos	R-Cos	U-Ent	Density	M-ADE	M-Optimal	M-FocusADE
ADE	0.0024	0.0224	0.0235	0.0352	0.0024	0.0992	0.2306	***0.3057***	0.1894
Dosage	0.8204	0.8569	0.8709	0.8692	0.8779	0.8691	0.9250	***0.9268***	0.9214
Drug	0.8809	0.9282	0.9302	0.9202	0.9282	0.9108	***0.9502***	0.9502	0.9471
Duration	0.0276	0.0276	0.0376	0.1276	0.0276	0.0514	0.7124	***0.7230***	0.6801
Form	0.5918	0.8324	0.8424	0.8354	0.8424	0.9059	***0.9077***	0.9051	0.9039
Frequency	0.7243	0.8493	0.8594	0.8403	0.8523	0.8981	0.8914	0.8832	***0.8915***
Reason	0.1855	0.3745	0.3768	0.3712	0.3794	0.2779	0.5527	***0.5689***	0.5583
Route	0.6375	0.7823	0.7856	0.7892	0.7819	0.8127	0.8582	0.8614	***0.8640***
Strength	0.8079	0.9112	0.9178	0.9124	0.9138	0.9057	***0.9609***	0.9599	0.9576
Micro F1	0.8293	0.9289	0.9282	0.9311	0.9286	0.9225	0.9413	***0.9484***	0.9349
Macro F1	0.5198	0.6205	0.6271	0.6334	0.6229	0.6368	0.7766	***0.7871***	0.7681

Table 1: F1 scores for baseline, non-meta, and meta-model strategies. Density ranks sentences by density features (f_5+f_6) only (no meta-model). Bold-italic marks the best value in each row. Per-tag scores are from a representative seed (42). Macro F1 is the unweighted mean over the nine entity types.

Model	Lenient Micro-F1	Annotated Sentences
GatorTron-MRC	0.9506	7,955
BERT-MIMIC-MRC	0.9489	7,955
M-Optimal (Ours)	**0.9475**	**5,256**
RoBERTa-MIMIC-MRC	0.9465	7,955
BERT-MRC	0.9440	7,955

Table 2: Top-5 results on the 2018 n2c2 ADE benchmark under lenient evaluation. Baseline scores are taken from Peng et al. (2023) (their Table 1) and are trained on the full training set. Our active learning setting uses fewer annotated sentences while achieving a micro-F1 of 0.9475 (mean over five seeds) and a macro-F1 of 0.7871 (unweighted across the nine entity types).

gains in held-out F1. This explains both the rare-type improvements and the macro-level gains at a fixed budget: acquisition-behaviour diagnostics show batches enriched with ADE, Reason, and Duration spans (Table 4), which mirrors the per-entity improvements and the higher Macro F1 (Table 1). The aggregate trends are stable across five seeds under identical training and evaluation settings, and the efficiency claim is supported by achieving comparable benchmark performance with fewer annotated sentences than prior systems trained on the full dataset (Table 2). Remaining errors are dominated by span boundary ambiguity in adverse events and lexical overlap between indication-like phrases and medication mentions; the full confusion matrix and analysis are provided in Appendix E.

See Section 8 for scope and threats; immediate extensions include matched multi-seed non-meta 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-neighbour search.

7 Conclusion

We introduced a meta-model-driven active learning framework for clinical NER that predicts a proxy performance gain (PPG) for each unlabelled sentence by integrating uncertainty, semantic diversity, ontology alignment, and an MWOM cue. The selector is trained using gold PPG labels measured as validation ΔF1 during simulated acquisitions, enabling it to prioritise sentences with the highest expected contribution to downstream performance.

Across five random seeds, M-Optimal achieved a mean Micro F1 of 0.9475 (SD 0.0004; 95% CI [0.9469, 0.9480]) while annotating 1,250 of 3,949 pool sentences (32%). Counting the initial training set and the acquired pool, the final model uses 5,256 labelled sentences, which is 34% fewer than the full 7,955-sentence training set used by standard supervised baselines. In the representative run, Macro F1 reached 0.7871, exceeding the strongest non-meta baseline (0.6334), with particularly large gains for low-prevalence types such as 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 reduce boundary ambiguity, 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 assess 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.

True-cost reporting is partial. Due to time constraints, we report true-cost diagnostics (span/token counts) for M-Optimal only; a fully cost-normalised comparison (equal span/token budgets across all strategies) is left to future work.

LLM-based active learning comparisons. Recent work explores using LLMs not only for querying but also for generation and low-cost annotation within the active learning loop (Xia et al., 2025). We do not include a direct comparison to LLM-based active learning frameworks, nor do we measure how LLM-assisted selection/annotation would interact with PPG learning. Establishing such comparisons and cross-dataset generalisability is an important direction for future work.

Model family and labelling formalism. We instantiate the learner as an encoder-only BioClinicalBERT-CRF with BIO tagging. While the PPG-based querying objective is model-agnostic, extending to encoder-decoder or decoder-only formulations (e.g., text-to-text structured prediction such as TANL; Paolini et al., 2021) would require deriving uncertainty/confidence cues from generation scores (e.g., token entropies or sequence log-probabilities) and aligning cost/coverage accounting with generated spans. A systematic benchmark across model families and label formalisms is left to future work.

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

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.

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.

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*.

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

Jiaxing Liu and Zoie S. Y. Wong. 2024. Utilizing active learning strategies in machine-assisted annotation for clinical named entity recognition: a comprehensive analysis considering annotation costs and target effectiveness. *Journal of the American Medical Informatics Association*, 31(11):2632–2640.

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*.

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.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations (ICLR)*. ArXiv:2101.05779.

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.

Cheng Peng, Xi Yang, Zehao Yu, Jiang Bian, William R Hogan, and Yonghui Wu. 2023. Clinical concept and relation extraction using prompt-based machine reading comprehension. *Journal of the American Medical Informatics Association*, 30(9):1486–1493.

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.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers.

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.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

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.

Robert Vacareanu, Enrique Noriega-Atala, Gus Hahn-Powell, Marco A. Valenzuela-Escarcega, and Mihai Surdeanu. 2024. Active learning design choices for ner with transformers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 321–334.

Qunbo Wang, Ruoyu Sun, Lin Xu, Tingwen Liu, and Guoliang Jin. 2024. BEAL: Deep active learning for multi-label text classification. *Scientific Reports*, 14:28246.

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. 2019. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322.

Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, Branislav Kveton, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Sungchul Kim, and 15 others. 2025. From selection to generation: A survey of LLM-based active learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14552–14569, Vienna, Austria. Association for Computational Linguistics.

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.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A survey on active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190. Association for Computational Linguistics.

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.

B Meta-model-driven AL pipeline

B.1 Acquisition behaviour: rare-span coverage

We counted gold spans of ADE, Reason, and Duration in each queried batch using B-tag occurrences (each B-tag marks a new span; I-tags are treated as continuations of the same span). Table 4 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). Residual confusions are concentrated in boundary cases for adverse events and in overlaps between indication-like phrases and medication mentions; the full confusion matrix and discussion are provided in Appendix E.

C Annotation cost analysis

Annotation cost (true effort). Because sentence counts can hide substantial variation in annotation effort, we additionally report the gold-label annotation cost of the M-Optimal queried batches under the fixed acquisition schedule (k=250 in Rounds 1–3 and k=500 in Round 4). We quantify cost as (i) the number of entity spans (all nine types; BIO spans counted from gold labels) and (ii) the number of entity tokens (gold tokens inside any entity span), alongside the total number of tokens for context (Table 5). This directly addresses the concern that improvements under a fixed sentence budget could be driven by selecting systematically denser sentences (e.g., via f_5/f_6), which increases human effort even when k is held constant. Due to time constraints, we report these true-cost diagnostics for M-Optimal only; a fully cost-normalised comparison (equal span/token budgets across all strategies) is left for future work.

D Comparison of Regression Models

To support the selection of Random Forest as the meta-regressor, Table 6 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.

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

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 3: 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.

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

Round	k (nominal)	Entity spans	Entity tokens	Total tokens	Cum. k
1	250	238.0	805.1	25,894.6	250
2	250	78.3	221.2	23,996.8	500
3	250	87.9	257.2	24,232.4	750
4	500	173.6	575.6	50,271.7	1,250

Table 5: True annotation cost for M-OPTIMAL (gold labels), reported as entity spans and entity tokens. Counts follow the fixed AL schedule (k=250 in Rounds 1–3; k=500 in Round 4).

E 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.

F Additional Results: Performance by Entity

Figure 3 summarises aggregate performance across entity types, reporting the mean F1 for each strategy averaged over tags. Overall, meta-model-driven methods achieve higher average F1; M-Optimal attains the strongest mean performance, consistent with incorporating additional meta-features (including MWOM) that capture uncertainty, representation-level signals, and domain-specific cues. The error bars (standard deviation across entity types) indicate that performance varies substantially between tags, reflecting heterogeneous difficulty across the label set.

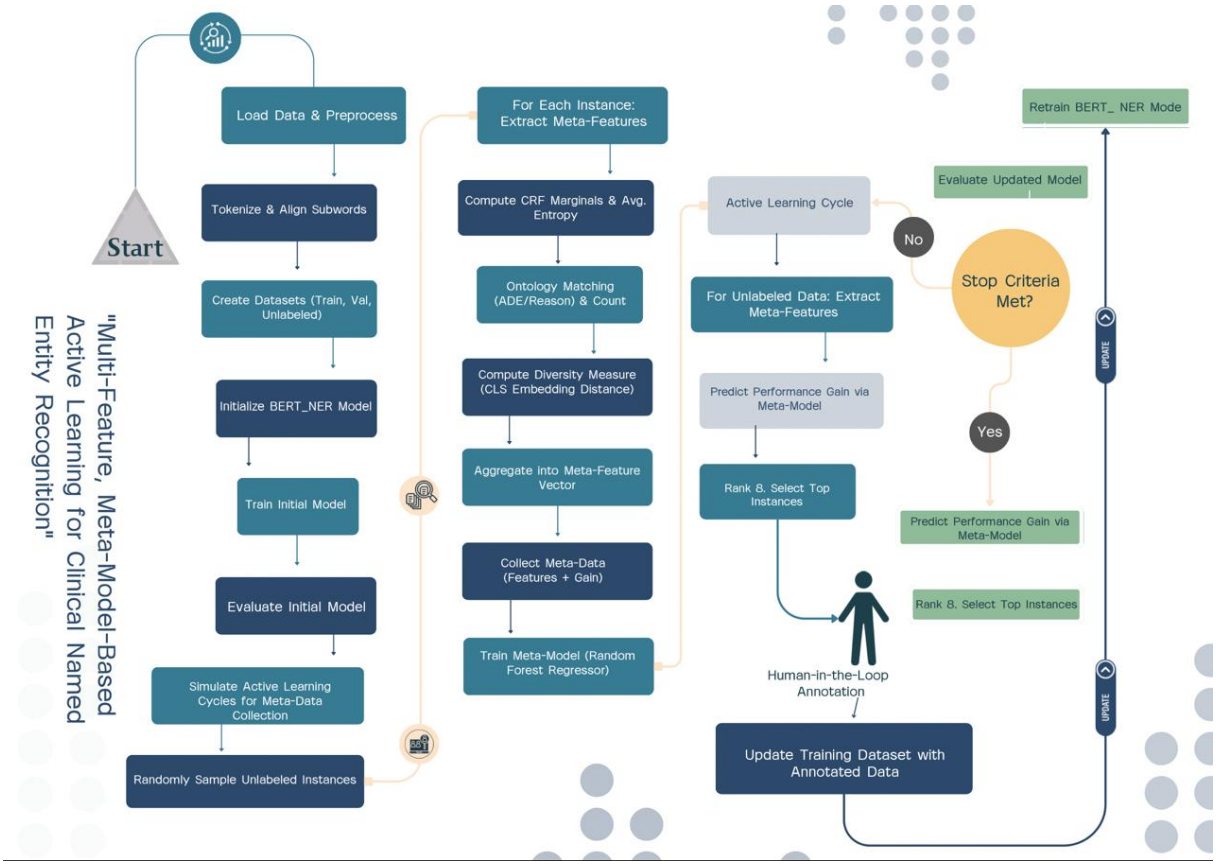

Figure 1: Overview of the meta-model-driven AL pipeline. The figure outlines the process of meta-feature extraction, proxy gain computation, and integration into the AL loop. **Human annotation steps are simulated by programmatically retrieving ground truth labels from the held-out CLiPS subset, ensuring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

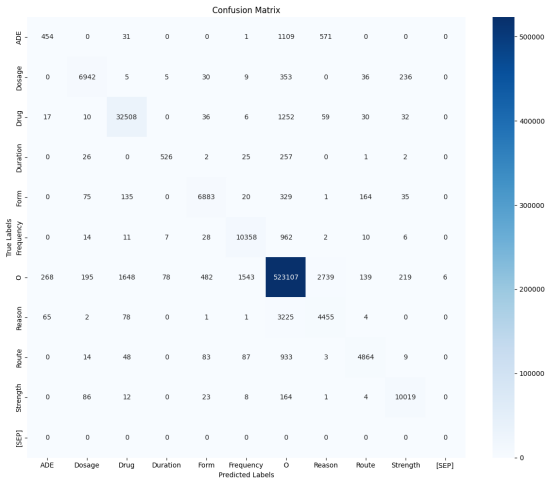

Figure 2: Confusion matrix after the final AL cycle.

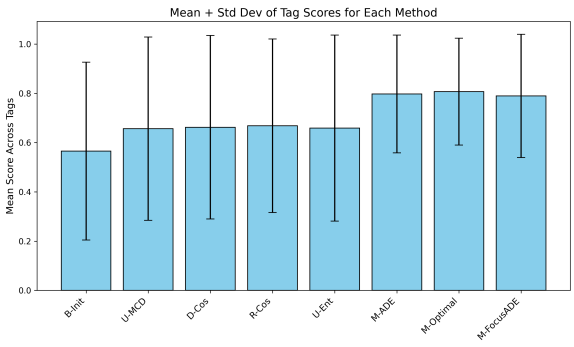

Figure 3: Mean F1 across entity types for each strategy. Error bars denote the standard deviation of per-entity-type F1 scores.

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	**0.9484**	0.9397
Macro F1	**0.7871**	0.7660

Table 6: Comparison of F1-scores between Random Forest (M-Optimal) and XGBRegressor across entity types. Random Forest shows stronger performance on several low-prevalence types (e.g., ADE, Duration), while XGBRegressor is higher on Frequency.