A Herd of Language Models Makes a Better Zero-shot Annotator for Clinical Named Entity Recognition

Seiji Shimizu, Shoko Wakamiya, Eiji Aramaki
Nara Institute of Science and Technology (NAIST)
shimizu.seiji.so8@is.naist.jp

Abstract

Clinical named entity recognition (NER) remains difficult to scale due to the high cost of manual annotation. Although large language models (LLMs) enable zero-shot annotation, their performance on clinical NER is still limited. To this end, we improve the annotation quality by aggregating annotations from *a herd of diverse LLMs*, including general-purpose, medically adapted, and NER-specialized models. A key challenge in this multi-LLM setting is effectively leveraging entities extracted by only a minority of models: although they account for a substantial portion of true positives, they are heavily intermixed with noise. To address this, we introduce **MARY**, a label-modeling method for **M**ulti-LLM **A**nnotation using **R**epresentation learning to capture contextual similarit**Y**. During aggregation, MARY selectively incorporates minority-extracted entities whose contexts are similar to those of majority-extracted entities, yielding more reliable and comprehensive annotations. Experimental results show that MARY improves the average F1 score by 8.6% over vanilla zero-shot baselines while reducing annotation costs[1].

1 Introduction

Clinical named entity recognition (NER) is a core task in clinical NLP (Uzuner et al., 2011; Stubbs and Özlem Uzuner, 2015; Henry et al., 2020; Fraile Navarro et al., 2023), supporting downstream applications such as cohort identification (Lopez et al., 2025), clinical decision support (Kocaman et al., 2025), and adverse event detection (Yanagisawa et al., 2025). However, building high-performing NER systems remains challenging due to the difficulty in obtaining annotated clinical corpora (Li et al., 2023). Annotated data for the target NER task is often not readily accessible due to strict data-sharing restrictions (Laparra et al., 2020). Moreover, expert annotation is costly and time-consuming, making large-scale manual annotation impractical (Chapman et al., 2011). Therefore, developing effective automatic annotation methods is critical for the scalable development of high-performing clinical NER systems.

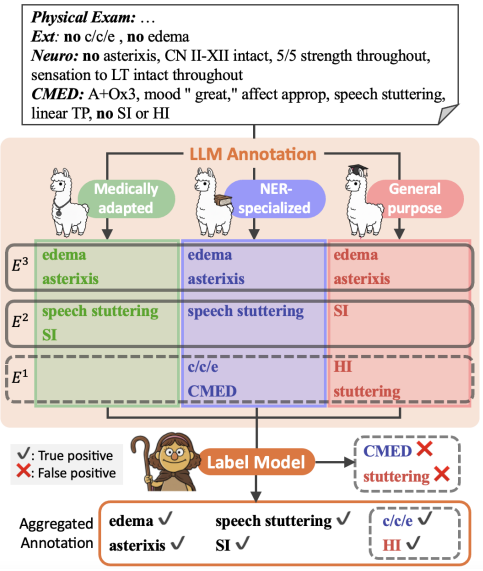

Figure 1: Overview of our zero-shot annotation approach with [PROBLEM] (medical problem) extraction shown as an example. Clinical notes are annotated by *a herd of diverse LLMs*, including general-purpose, medically adapted, and NER-specialized models. This multi-LLM setup produces sets of extracted entities with different levels of model agreement $\{E^w\}_{w=1}^{3}$, where w denotes the number of models that extract a given entity. Our label model aggregates entities across all agreement levels and selectively incorporates entities from E^1 when they are contextually consistent with majority-extracted entities (e.g., "*Section header*: no + [PROBLEM]").

Recently, large language models (LLMs) have been increasingly utilized as annotators to train much smaller, computationally efficient NER models (Hsu and Roberts, 2025; Hu et al., 2026; Xiao et al., 2023). Zero-shot in-context learning (ICL)

[1] https://github.com/seiji-shimizu/MARY_ClinicalNER

Findings of the Association for Computational Linguistics: ACL 2026, pages 12327–12344
July 2-7, 2026

can offer a scalable alternative to manual annotation, enabling flexible entity extraction across diverse target tasks (Hu et al., 2024; del Moral-González et al., 2025). Nevertheless, even state-of-the-art LLMs with tens of billions of parameters exhibit limited zero-shot performance in clinical NER (Naguib et al., 2024; Hu et al., 2026; Chen et al., 2025a). Although medical domain adaptation (Christophe et al., 2024) and NER specialization (Zhou et al., 2024) can yield gains in in-domain performance, these methods often come at the cost of reduced robustness on unseen data (Dorfner et al., 2024). Consequently, individual off-the-shelf LLMs remain insufficient as reliable zero-shot annotators for clinical NER.

To address the limited performance of individual models, we improve annotation quality by aggregating outputs from *a herd of diverse LLMs*, including general-purpose (Grattafiori et al., 2024), medically adapted (Christophe et al., 2024), and NER-specialized LLMs (Zhou et al., 2024). An overview of our approach is illustrated in Fig. 1. For instance, when extracting [PROBLEM] (medical problem), the medically adapted model provides high-precision annotations, while the other models offer noisier but broader coverage. When aggregated effectively, their outputs provide more reliable and comprehensive annotations than any single model alone.

Although seemingly straightforward, achieving effective aggregation is challenging. As illustrated in Fig. 1, a multi-LLM setup produces sets of extracted entities with varying levels of model agreement $\{E^w\}_{w=1}^3$, where w denotes the number of models that extract a given entity. Our experiments show that **(i)** entities extracted by **a majority of models** ($E^2 \cup E^3$) are true positives with high precision, but relying solely on these entities leads to limited annotation coverage; and **(ii)** incorporating entities extracted by only **a minority** (E^1) can greatly improve recall, but these entities are **heavily intermixed with noise** (see Sect. 4.4). For example, the abbreviations "c/c/e" and "CMED" are extracted only by the NER-specialized model. Based on the surrounding context, "CMED" is likely noise (a section header) rather than a clinical entity. In contrast, "c/c/e" is likely a [PROBLEM], as it appears in a context similar to majority-extracted entities, such as "edema" and "asterixis," following the pattern "*Section header*: no + [PROBLEM]". Selectively incorporating such **contextually consistent minority-extracted entities** is essential for expanding annotation coverage beyond simple majority voting while maintaining high precision.

To tackle this challenge, we introduce MARY, a label-modeling method for **M**ulti-LLM **A**nnotation using **R**epresentation learning to capture contextual similarit**Y** (Fig. 2). MARY promotes the integration of entities from E^1 **when they appear in similar contexts to majority-extracted entities**. Specifically, it employs domain-adaptive pretraining and contrastive learning to align entity representations with class-specific centroids estimated from E^3. By clustering true positives in the representation space, MARY effectively incorporates contextually consistent minority entities while isolating contextually dissimilar noise.

We evaluate MARY on diverse target tasks derived from the i2b2 2010 (Uzuner et al., 2011), i2b2 2014 (Stubbs and Özlem Uzuner, 2015), and n2c2 2018 (Henry et al., 2020). Across evaluation settings, MARY consistently outperforms both individual models and label-model baselines including majority voting. When combined with zero-shot refinement methods, MARY yields an average *+8.6* F1 improvement over Llama-3.3-70B-instruct. Furthermore, our analysis demonstrates that MARY selectively incorporates true-positive entities from E^1 while reducing the manual annotation burden.

Our contributions are as follows:

- We explore a *multi-LLM annotation framework for zero-shot clinical NER*, aggregating outputs from general-purpose, medically adapted, and NER-specialized models.
- We propose a novel label-modeling method that effectively incorporates minority-extracted entities by capturing their contextual similarity to majority-extracted entities.
- We conduct extensive experiments showing that our approach consistently outperforms single-model annotation, yielding average F1 gains of up to *8.6* points over vanilla zero-shot baselines.

2 Related Work

2.1 Zero-shot NER

LLMs have recently been applied to clinical NER with in-context learning (ICL) (Agrawal et al., 2022; Averly and Ning, 2025). However, several studies have reported their underperformance especially compared to fully-supervised method (Naguib et al., 2024; Hu et al., 2026; Chen et al., 2025a). Prompt engineering has been widely explored to mitigate this limitation (Hu et al., 2024;

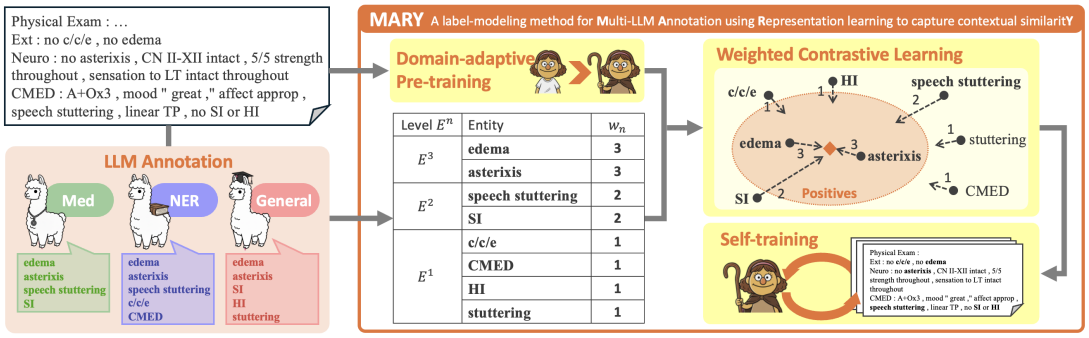

Figure 2: Overview of LLM annotation and MARY. Clinical notes are annotated using three LLMs. MARY aggregates these annotations through: (i) **domain-adaptive pre-training** to encode contextual similarity, (ii) **weighted contrastive learning** based on model agreement w_n, and (iii) **self-training** to expand entity coverage. In (ii), we first compute a centroid (orange diamond) from the average embeddings of entities in E^3. All entities are then pulled toward this centroid with a force proportional to their w_n. This allows minority-extracted entities in E^1 (e.g., "c/c/e") to be recognized as positives **only** when they are contextually similar to E^3, while ensuring majority-extracted entities in E^2 (e.g., "speech stuttering") remain positive **even** when they are contextually dissimilar.

Sivarajkumar et al., 2024), but performance remains highly sensitive to prompt design (Ceballos-Arroyo et al., 2024), limiting its practical reliability.

Beyond prompt engineering, some work has explored task-specific fine-tuning (Jiao et al., 2023; Zhou et al., 2024) to mitigate the lack of NER-specific supervision signals in instruction-tuning (Wang et al., 2022). Zhou et al. (2024) employs targeted distillation methods, fine-tuning lightweight LLMs on synthetic NER datasets generated by ChatGPT. However, the resulting models often fail to generalize to real-world clinical notes that differ from their training distributions. Another line of work improves extraction performance by refining LLM outputs through multiple rounds of inference (Xie et al., 2024; Wang et al., 2025; Lu et al., 2025). While these approaches primarily focus on improving **stand-alone zero-shot performance**, our multi-LLM setup is orthogonal to such methods and can be used in conjunction with them.

2.2 Label Modeling

In weakly supervised learning (WS), noisy but easily obtained supervision signals are combined into high-quality labels using a *label model* (Zhang et al., 2022; Ratner et al., 2017). This paradigm reduces annotation cost and has been successfully applied in the clinical domain (Datta and Roberts, 2023; Wang et al., 2021; Cusick et al., 2021). Conventional WS approaches rely on human-defined labeling functions (LFs), such as dictionaries or data-programming rules, which are often labor-intensive to construct in the clinical domain (Hsu and Roberts, 2025).

More recently, ICL has been explored to reduce LF construction costs (Smith et al., 2024; Su et al., 2023), though its use in NER remains limited. A related line of work explores *LLM ensembles*, where predictions from multiple LLMs are integrated (Chen et al., 2025b; Ding et al., 2025). Xiao et al. (2025) apply this idea to zero-shot NER by aggregating outputs from *multiple general-purpose LLMs*, taking the union of extracted mentions and assigning entity types through majority voting. However, this approach does not explicitly address the noise introduced by aggregating entity mentions from diverse models. In contrast, MARY explicitly filters noise among minority-extracted entities by leveraging contextual similarity, substantially improving precision over union-based aggregation while maintaining high recall.

3 Method

Fig. 2 illustrates an overview of our zero-shot annotation framework.[2] We first employ a diverse set of LLMs to annotate unlabeled data, producing multi-LLM annotations (Sect. 3.1). Although entities extracted by a majority of models are typically high-precision true positives, relying on them alone yields limited annotation coverage. Incorporating entities extracted by only a minority of models is therefore essential for improving recall, despite their higher noise levels. To address this challenge, we introduce **MARY**, a label-modeling framework that selectively incorporates minority-extracted entities by leveraging their contextual similarity to majority-extracted entities (Sect. 3.2).

[2]The algorithm for MARY is detailed in Appendix D.4.

3.1 LLM Annotation

We denote an unlabeled data as D, which serves as the target for automatic annotation. Each sample $d \in D$ is annotated by a set of m LLMs denoted $\{M_i\}_{i=1}^{m}$, where $m = 3$ in our experiments. In a zero-shot setting, each LLM produces a set of entity mentions with their corresponding entity types:

$$M_i(d) = \{(e_n, c_n)\}_{n=1}^{N},$$

where N is the total number of extracted entities by model M_i. To train the label model with a sequence labeling objective, we convert each model's entity-level output into BIO-tagged sequences using a rule-based resolver $\mathcal{R}$:

$$\mathcal{R}(d, M_i(d)) = \mathbf{y}.$$

These vanilla annotations may optionally be refined using existing advanced zero-shot methods (Wang et al., 2025; Xie et al., 2024; Lu et al., 2025). While MARY is compatible with such refinement strategies, we focus on vanilla zero-shot predictions in this section for clarity.

3.2 MARY

Our goal is to aggregate the multi-LLM annotations $\{M_1(d), \ldots, M_m(d)\}$ into a single annotation using a label model L. We define L as a composite architecture consisting of an encoder-based backbone G (e.g., RoBERTa-base) and a sequence labeling head F. Specifically, for a given input d, the model produces the final annotation $\hat{\mathbf{y}}$ as follows:

$$\hat{\mathbf{y}} = L(d) = F(G(d))$$

3.2.1 Domain-Adaptive Pre-training

To encourage the encoder G to capture contextual similarity among clinical entities, we perform domain-adaptive pre-training on D using a masked language modeling objective (Devlin et al., 2019). This step helps entities appearing in similar clinical contexts obtain similar representations. We initialize G with these domain-adapted weights.

3.2.2 Weighted Contrastive Learning

We employ contrastive learning weighted by w_n, where w_n denotes the number of LLMs that extracted an entity e_n. Let $G(d)$ represent the contextualized token representations for d. Each entity mention e_n is represented by an embedding vector $\mathbf{e}_n$[3]. To anchor entity representations, we first compute class centroids μ_c using entities with a full model agreement ($w_n = m$), providing high-precision anchors within each sample d. For each entity type c, the centroid is computed as:

$$\mu_c = \frac{1}{|E_c^m|} \sum_{\mathbf{e}_n \in E_c^m} \mathbf{e}_n,$$

where E_c^m denotes a set of entity embeddings that are (i) extracted by all m models and (ii) assigned to entity type c.

We then perform contrastive learning using the union of all extracted entities. Because domain-adaptive pre-training already places contextually similar entities close in embedding space, **the weighting behaves intuitively**: (i) Entities with **low** w_n (minority-extracted) are pulled weakly and reach the centroid **only** when they are contextually similar to entities with $w_n = m$. (ii) Entities with **high** w_n (majority-extracted) are pulled strongly and end up close to the centroid, **even** if they are not contextually similar. Conversely, minority-extracted noise (entities with low w_n and dissimilar contexts) remains distant from the centroid and is effectively filtered out. The weighted contrastive loss is defined as:

$$\mathcal{L}_{\text{ct}} = -\frac{1}{N} \sum_{n=1}^{N} w_n \log \frac{\exp(\text{sim}(\mathbf{e}_n, \mu_{c_n}))}{\sum_{c=1}^{C} \exp(\text{sim}(\mathbf{e}_n, \mu_c))},$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.

Parallel to the contrastive objective, the classification head F is trained using a standard cross-entropy loss over the BIO sequences $\mathbf{y}$ resolved from the LLM annotations:

$$\mathcal{L}_{\text{ce}} = \text{CrossEntropy}(L(d), \mathbf{y}).$$

The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{ct}} + \mathcal{L}_{\text{ce}}.$$

Using this objective, we train the label model L on the complete set of multi-LLM annotations $\{M_1(d), \ldots, M_m(d)\}$.

3.2.3 Self-training

After training with weighted contrastive learning, we further enhance the recall of the label model L through self-training. While weighted contrastive learning improves the encoder G's ability to cluster semantically similar entities, it may still produce conservative predictions since the predictions are

[3]While an entity mention is formally a token span $[t_{start}, \ldots, t_{end}]$, we denote its representation as a single vector for simplicity.

anchored by class centroids. By training on its own predictions, self-training helps the model to "bootstrap" itself and reinforce entity patterns across D.

For each sample d, we obtain the model's current BIO sequence prediction $\tilde{\mathbf{y}} = L(d)$. We fine-tune the entire label model using the standard cross-entropy objective:

$$\mathcal{L}_{\text{st}} = \text{CrossEntropy}(L(d), \tilde{\mathbf{y}}).$$

4 Experiments

4.1 Setup

For multi-LLM annotation, we employ Med42-8B[4], UniNER-7B-type[5], and Llama-3.3-70B[6], and use their vanilla zero-shot performances as baselines. We also include the following **zero-shot annotation methods** to refine each model's annotation.

Self-consistency (Xie et al., 2024): Each sample is annotated multiple times by each LLM, and only entities extracted more than a predefined threshold are retained to denoise model outputs.

Self-verification (Wang et al., 2025): For each extracted entity, the LLM is asked to verify whether the prediction is correct, and only confirmed entities are retained.

Conflict resolution: This method was originally proposed in a multi-agent setting using a single backbone model (Lu et al., 2025). We extend this method to the multi-LLM setup. Specifically, when two LLMs produce conflicting annotations, we verify the annotations using Llama-3.3-70B.

We also compare MARY with the following **baseline label models** commonly used in weakly supervised NER (Zhang et al., 2021).

Majority Voting (MV) (Lison et al., 2020): First, select entity mentions extracted by at least T models (i.e., entities in $\bigcup_{w=T}^{m} E^w$), and then apply majority voting to determine their entity types.

Hidden Markov Model (HMM) (Lison et al., 2020): This model treats true BIO tags as latent states and the outputs of multiple labeling functions as observed emissions. It estimates tag transitions and labeling-function emission probabilities to infer the most probable sequence of BIO labels.

Conditional HMM (CHMM) (Li et al., 2021): This model extends HMM by conditioning transition and emission probabilities on contextual embeddings of encoder-based models, enabling more expressive context-aware label inference.

Evaluation includes combinations of these zero-shot annotation methods and label models. **This setup allows us to assess whether MARY (1)** improves over single-model zero-shot methods, **(2)** outperforms baseline label models, and **(3)** is compatible with these zero-shot annotation methods.

4.2 Datasets

We evaluate on three clinical NER datasets with diverse entity types: **i2b2 2010** (Uzuner et al., 2011), containing *problem*, *treatment*, and *test* entities; **i2b2 2014** (Stubbs and Özlem Uzuner, 2015), focusing on the extraction of protected health information (PHI), including patient demographics and identifiers; and **n2c2 2018** (Henry et al., 2020), which targets adverse drug events and drug-related entities such as drug names, preprocessing details are provided in Appendix C.

4.3 Implementation Details

All experiments were conducted on two NVIDIA A100 GPUs (80GB). Llama-3.3-70B is quantized to 4-bit precision for efficient inference. For self-consistency and self-verification, we re-implement the methods following Xie et al. (2024); Wang et al. (2025). For conflict resolution, we extend the multi-agent procedure from Lu et al. (2025) and resolve conflicts at both mention and type levels. Implementations of MV, HMM, and CHMM are adapted from publicly available resources from the original publications (Lison et al., 2020; Li et al., 2021). For MARY, we use RoBERTa-base[7] as G and use a fixed set of hyperparameters across all datasets, assuming no manually annotated development set is available. The hyperparameters, prompt templates, and additional implementation details are provided in Appendices D and H.

4.4 Results

Table 1 summarizes the performance of MARY and the baselines in micro-averaged F1 score, precision (Prec.), and recall (Rec.). We also include a few-shot baseline from Liu et al. (2022), with implementation details provided in Appendix D.2. **Throughout all experiments, results are averaged over three different runs with different seeds.**

[4] https://huggingface.co/m42-health/Llama3-Med42-8B

[5] https://huggingface.co/UniNER-7B-type/UniNER-7B-type

[6] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[7] https://huggingface.co/FacebookAI/roberta-base

Annotation Method	Model	i2b2 2010			i2b2 2014			n2c2 2018			Avg. F1
		F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	
Few-shot	Llama-3.3-70B	59.0	57.3	60.7	58.5	57.9	59.1	49.2	40.9	61.6	55.6
Vanilla zero-shot	Med42-8B	42.7	53.4	35.6	27.7	20.7	41.5	43.0	35.9	53.8	37.8
	UniNER-7B-type	43.9	51.5	38.2	33.2	38.7	29.1	40.7	40.2	41.2	39.3
	Llama-3.3-70B	53.0	52.6	53.5	56.7	53.6	60.1	48.1	40.4	59.3	52.6
	MV (T=1)	53.6	47.6	**61.2**	35.1	23.5	**68.8**	41.1	31.7	58.3	43.3
	MV (T=2)	52.3	62.4	45.0	51.8	59.5	45.9	48.9	46.6	51.5	51.0
	MV (T=3)	34.6	**68.6**	23.1	25.9	**84.7**	15.3	48.6	**61.1**	40.3	36.4
	HMM	51.7	51.8	51.6	27.7	18.0	59.6	44.4	35.4	**59.6**	41.3
	CHMM	52.4	51.3	53.6	32.4	22.2	60.0	46.4	37.6	60.5	43.7
	MARY	**58.6**	61.1	56.2	**65.6**	67.0	64.4	**51.2**	45.9	57.9	**58.5**
Self-consistency	Med42-8B	44.6	59.0	35.9	30.8	24.1	42.4	41.8	37.8	46.7	40.3
	UniNER-7B-type	44.9	54.7	38.1	34.3	41.5	29.3	39.7	45.4	35.3	40.4
	Llama-3.3-70B	53.4	55.5	51.5	58.1	56.2	60.2	48.4	42.8	55.7	53.3
	MV (T=1)	54.6	49.4	**61.0**	38.1	26.4	**68.4**	40.3	31.6	55.6	44.3
	MV (T=2)	52.5	63.7	44.6	53.7	62.9	46.9	48.1	47.2	49.1	51.4
	MV (T=3)	33.3	**72.6**	21.6	27.7	**87.0**	16.5	47.1	**69.3**	35.7	36.0
	HMM	53.9	55.1	52.8	35.0	25.7	55.1	46.4	39.7	**55.9**	45.1
	CHMM	53.9	53.3	54.5	35.3	25.0	60.4	42.2	34.7	53.8	43.8
	MARY	**56.7**	60.7	53.1	**65.9**	68.7	63.4	**53.3**	58.4	49.0	**58.6**
Self-verification	Med42-8B	42.3	54.6	34.5	28.9	22.5	40.4	45.0	38.9	53.3	40.0
	UniNER-7B-type	43.7	56.2	35.7	32.9	42.8	26.8	39.8	47.0	34.5	39.9
	Llama-3.3-70B	49.0	55.5	43.8	60.7	62.1	59.4	48.9	42.4	57.8	52.9
	MV (T=1)	**53.9**	50.6	**57.6**	38.4	26.7	**68.0**	44.0	35.2	58.6	45.4
	MV (T=2)	48.6	63.7	39.3	52.8	65.0	44.4	51.1	50.4	51.7	50.8
	MV (T=3)	28.6	**69.0**	18.0	24.6	**89.2**	14.3	46.4	**64.6**	36.2	33.2
	HMM	52.2	56.0	49.0	35.1	25.8	54.9	49.0	42.8	57.2	45.4
	CHMM	52.5	54.1	51.0	35.7	25.4	60.0	48.1	40.3	**59.6**	45.4
	MARY	53.8	60.4	48.6	**65.6**	68.5	63.0	**53.9**	49.3	59.4	**57.8**
Conflict resolution	Med42-8B	45.2	62.6	35.4	41.6	43.0	40.2	48.8	44.9	53.5	46.1
	UniNER-7B-type	46.1	59.7	37.5	38.5	58.1	28.8	47.9	57.9	40.8	46.1
	Llama-3.3-70B	55.7	62.4	50.3	62.4	66.3	58.9	53.0	49.7	56.8	57.0
	MV (T=1)	58.3	56.9	**59.7**	54.4	45.7	**67.2**	49.2	42.6	58.4	54.0
	MV (T=2)	52.3	67.0	42.9	54.4	68.0	45.4	53.4	55.8	51.3	53.4
	MV (T=3)	32.2	**72.4**	20.7	25.9	**84.7**	15.3	49.6	**69.2**	38.7	35.9
	HMM	57.0	61.6	53.0	52.0	48.7	55.7	53.5	49.1	58.7	54.2
	CHMM	57.6	60.5	55.1	50.7	44.7	58.6	52.7	46.8	**60.3**	53.7
	MARY	**58.6**	63.8	54.1	**65.7**	69.7	62.2	**59.3**	60.5	58.3	**61.2**

Table 1: Performance in micro F1 score (F1), precision (Prec.), and recall (Rec.). For each zero-shot annotation method, we report the performance of single-model annotation, baseline label models (highlighted in light gray), and MARY. The best score within each zero-shot annotation method is shown in bold and across all zero-shot annotation methods is underlined.

(1) Comparison with Single-model Annotation. Even the strongest single-model annotation with Llama-3.3-70B achieves only moderate F1 scores across datasets. Aggregating annotations via MARY yields substantial gains. In the vanilla zero-shot setting, MARY improves over Llama-3.3-70B by *+5.9* F1 points on average, despite the other contributing models (Med42-8B and UniNER-7B-type) exhibiting weaker standalone performance. Among zero-shot refinement strategies, *conflict resolution* consistently provides the largest improvements. When combined with MARY, this configuration yields an overall *+8.6* F1 improvement over vanilla zero-shot annotation with Llama-3.3-70B.

(2) Comparison with Baseline Label Models. MARY achieves highest F1 scores across nearly all settings. Consistent with prior findings (Zhang et al., 2021), MV provides a robust conventional baseline. MV with a threshold of $T = 1$ (taking the union of all extracted mentions) often achieves the highest recall, while $T = 3$ yields the highest precision, highlighting a clear precision–recall trade-off. However, improvement of MV from single-model annotation in F1 is inconsistent. In contrast, MARY maintains a balanced precision-recall profile. Although HMM and CHMM occasionally improve over single-model annotation when noise is mitigated, their gains are inconsistent[8].

[8]We further discuss the performance of baseline label models in Appendix E.

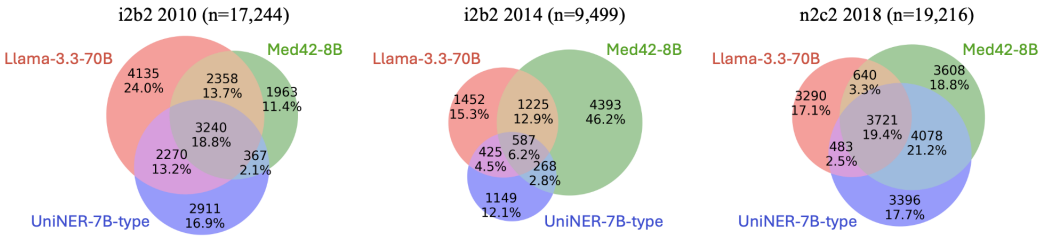

Figure 3: Venn diagrams of the entities extracted by the three LLMs. Here, n denotes the total number of entities extracted across all models, and percentages are computed relative to n. More than half of the entities are non-overlapping, highlighting the diversity of LLM annotations.

Setting	i2b2 2010	i2b2 2014	n2c2 2018	Avg.
w/o Med42-8B	55.7	**68.3**⋆	55.3⋆	59.8⋆
w/o UniNER-7B-type	**60.2**⋆	61.1	50.8	57.4⋆
w/o Llama-3.3-70B	57.9⋆	53.3	48.2	53.1⋆
All	58.6⋆	65.7⋆	**59.3**⋆	**61.2**⋆

Table 2: Ablation study of model contributions to MARY. We report the F1 score when removing each model from the multi-LLM setup (w/o). The ⋆ indicates an improvement over the strongest single-LLM baseline. While performance improves across configurations, utilizing all three LLMs yields the most robust results and the highest average F1 score.

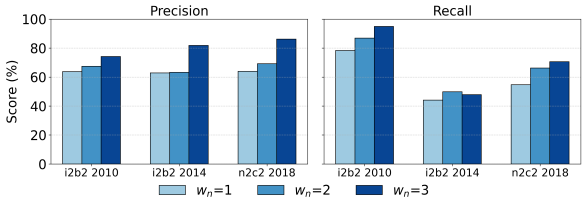

Figure 4: Binary classification performance of MARY on LLM-extracted entities, stratified by model agreement w_n. The left and right parts show precision and recall, respectively. MARY consistently identifies true-positive entities with high precision even with minimal model agreement ($w_n = 1$).

(3) Compatibility with Zero-shot Annotation Methods. When aggregated by MARY, annotation quality consistently improves over the best-performing single-model annotations across all zero-shot annotation methods. In particular, MARY successfully propagates the precision gains achieved through *conflict resolution* into the final aggregated output, indicating that it is compatible with a wide range of zero-shot annotation methods.

5 Discussion

In this section, we address the remaining research questions, denoted as **Q1** through **Q3**. For all analyses of MARY, we use *conflict resolution* as the base annotation method, as it is the best-performing setup in our evaluation.

Q1: Do the LLMs in our framework extract complementary subsets of clinical entities?

We analyze the overlaps among entities extracted by the three LLMs. As shown in Fig. 3, more than half of the entities in the union set are extracted by only a single model. Taking the union of all model outputs greatly increases annotation coverage and results in high recall, as reflected by MV ($T = 1$) in Table 1. Also, the intersection of all three models represents only a small fraction of the union. This suggests that the noise produced by each LLM is at least partially independent, as reflected by high precision of MV ($T = 3$).

The model-ablation shown in Table 2 confirms that aggregating annotations from all models yields the most robust and highest average F1 across datasets. Notably, even combining only two models already outperforms the stronger of the two on average, indicating that complementary entities are extracted even in small model subsets. Overall, these findings confirm that **the three LLMs provide complementary entity extractions, enabling more comprehensive annotation.**

Q2: Can MARY selectively incorporate entities extracted by only a minority of models?

To evaluate how MARY handles minority-extracted entities, we conduct a stratified analysis based on the level of model agreement $w_n \in \{1, 2, 3\}$. Within each stratum, we frame the task as a binary classification: for any entity originally extracted by an LLM, the prediction is 1 if it is retained in the final output of MARY, and 0 otherwise. This allows us to assess the performance of MARY in distinguishing true positive entities from noise across different levels of model agreement.

Fig. 4 summarizes the results. Across datasets, MARY identifies true positive entities extracted by only a single LLM ($w_n = 1$) with high pre-

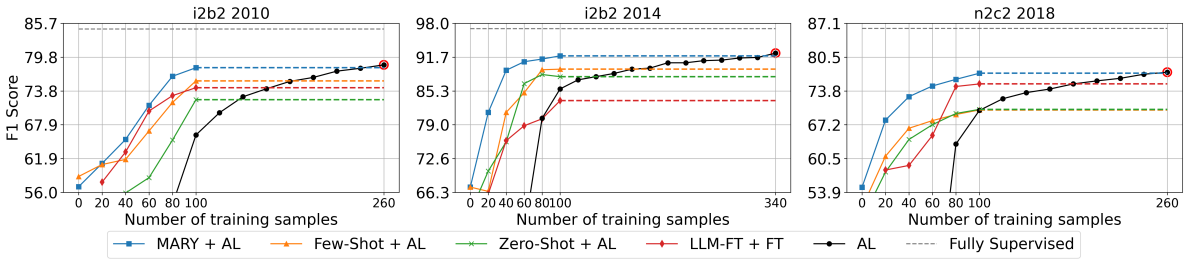

Figure 5: Performance of downstream NER models in low-resource fine-tuning settings. The upper dashed line denotes fully supervised training using all manual annotations. Across all datasets, MARY+AL consistently achieves the strongest performance at every annotation budget > 0. Notably, the performance attained by MARY+AL with 100 training samples is only matched by vanilla AL after 2.6 to 3.4 times more annotations.

cision. This results in higher recall than the MV ($T = 2$) baseline, while maintaining precision at a comparable level (further analysis is provided in Appendix E). We note that recall is lower on the i2b2 2014 dataset, primarily due to partial span predictions (e.g., "04" instead of "04/03/83"), resulting in strict-match misses. However, we demonstrate in the following section that such boundary errors can be corrected with minimal manual supervision. Overall, these findings confirm that **MARY identifies true positive entities overlooked by the majority of LLMs with high precision.**

Q3: Does MARY reduce manual annotation effort for downstream fine-tuning?

We evaluate MARY in a realistic **low-resource fine-tuning** scenario, where only a limited budget of manual annotations (up to 100 samples) is available. Using a combination of LLM annotations and the manually annotated data, we fine-tune a RoBERTa-based encoder with a token classification head as the NER model. As a strong baseline, we employ active learning (AL) from Su et al. (2022), which iteratively selects the most informative annotated samples based on model uncertainty. We compare the following fine-tuning strategies:

Zero-shot+AL: warm-starts AL using zero-shot annotations by Llama-3.3-70B.

Few-shot+AL: warm-starts AL using few-shot annotations by Llama-3.3-70B.

LLM-FT+FT (Hsu and Roberts, 2025): fine-tunes an LLM on manually annotated data, uses it to annotate the full training set, and then fine-tunes the NER model on the combined data.

MARY+AL: fine-tunes the NER model using both manual and MARY annotations and jointly updates MARY and the NER model at each AL iteration.

For AL, we acquire 20 manually labeled samples per iteration for a total of five iterations (100 samples). We also report the performance of a fully supervised model trained on all available annotations as an upper bound. Additional implementation details are provided in Appendix D.5.

Fig. 5 shows that all LLM-assisted methods substantially outperform vanilla AL, demonstrating the effectiveness of LLM annotation in low-resource settings. Across all datasets, MARY+AL consistently achieves the strongest performance under the same annotation budget. In particular, vanilla AL requires approximately *2.6 to 3.4* times more annotated samples to match the performance level that MARY+AL reaches with only 100 samples. These results indicate that **MARY is highly effective in low-resource settings, substantially reducing the cost of manual annotation.**

6 Conclusion

This paper explores aggregating annotations from *a herd of diverse LLMs*, including general, medically-adapted, and NER-specialized LLMs for zero-shot clinical NER. To effectively incorporate true positives overlooked by the majority of LLMs, we proposed MARY, a novel label-modeling method that leverages contextual similarity to denoise minority-extracted entities. Our experiments demonstrate that MARY improves zero-shot F1 scores by up to 8.6 points, consistently outperforming single- and label-model baselines. Furthermore, downstream fine-tuning results show that MARY reduces manual annotation effort by a factor of 2.6 to 3.4. Overall, this work demonstrates that leveraging model diversity provides a scalable path toward high-quality clinical NER annotation without the need for extensive human annotation.

Limitations

Despite the annotation quality improvement and fine-tuning effectiveness achieved by MARY, several limitations remain. **First**, our primary experiments were conducted using a fixed set of three LLMs. To this end, we explored alternative combinations in Appendix A, including Llama-based models of varying sizes (7B, 8B, and 70B) and diverse architectures (Qwe3, Phi-4, and Llama-3.3). While MARY consistently improved performance across all tested configurations, the original configuration of general-purpose, medically adapted, and NER-specialized models proved most effective. Future work should investigate (i) systematic strategies for identifying optimal model combinations for specific task and (ii) the scaling laws associated with increasing the number of models.

Second, since we operate in a zero-shot setting where an annotated validation set is unavailable, MARY relies on fixed hyperparameters. We demonstrate in Appendix B that the components of MARY, namely, DAPT, weighted contrastive learning, and self-training, ensures stable performance without the need for hyperparameter tuning. However, performance could potentially be further enhanced through hyperparameter optimization. Future work could explore unsupervised hyperparameter optimization or meta-learning techniques to adapt these parameters automatically in the absence of labeled data.

Third, the use of multiple LLMs introduces higher computational overhead and inference costs compared to single-model annotation. We have demonstrated that MARY can significantly improve the performance of a high-capacity model (70B) by aggregating it with much smaller, specialized models (7B and 8B). Nevertheless, the cumulative computational cost remains higher than querying a single large-scale model. Future research should explore active learning or uncertainty-based sampling methods to identify which samples truly require multi-LLM annotation, thereby reducing the computational burden while maintaining high annotation quality.

Finally, while our work focuses on the zero-shot annotation of clinical notes, other strategies, such as synthetic data generation for data augmentation, can also alleviate annotation burdens. However, we argue that zero-shot annotation of real-world clinical records is orthogonal to synthetic data generation. Due to strict data-sharing restrictions and the prevalence of institution-specific jargon (e.g., localized abbreviations and formatting), generating institution-specific synthetic notes is difficult with off-the-shelf LLMs. Conversely, zero-shot annotation utilizes existing, authentic documentation, thereby preserving the unique linguistic characteristics and specialized vocabulary inherent to specific clinical institutions. Future work should investigate the potential synergies between synthetic data augmentation and zero-shot annotation frameworks.

Ethical Considerations

The datasets used in this study are publicly available. Access to the i2b2 and n2c2 datasets requires the completion of appropriate credentialing and the signing of a data use agreement. All data usage complies with ethical standards. The proposed framework is intended to assist, not replace, human annotation, and any real-world deployment should include appropriate human evaluation.

Acknowledgments

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425 and JST, CREST Grant Number JPMJCR22N1, Japan.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Reza Averly and Xia Ning. 2025. Entity decomposition with filtering: A zero-shot clinical named entity recognition framework. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2935–2951, Albuquerque, New Mexico. Association for Computational Linguistics.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. 2024. Open (clinical) LLMs are sensitive to instruction phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71, Bangkok, Thailand. Association for Computational Linguistics.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K

Savova, and Ozlem Uzuner. 2011. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, Vipina K. Keloth, Kalpana Raja, Jimin Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A. Adelman, and 2 others. 2025a. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1):3280.

Zhijun Chen, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Ming Li, Likang Xiao, Dingqi Yang, Yikun Ban, Hailong Sun, and Philip S Yu. 2025b. Harnessing Multiple Large Language Models: A Survey on LLM Ensemble. *Preprint*, arXiv:2502.18036.

Clément Christophe, Praveenkumar Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al Mahrooqi, Avani Gupta, Muhammad Umar Salman, Marco AF Pimentel, Shadab Khan, and Boulbaba Ben Amor. 2024. Med42–evaluating fine-tuning strategies for medical LLMs: full-parameter vs. parameter-efficient approaches. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Marika Cusick, Prakash Adekkanattu, Thomas R. Campion, Evan T. Sholle, Annie Myers, Samprit Banerjee, George Alexopoulos, Yanshan Wang, and Jyotishman Pathak. 2021. Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. *Journal of Psychiatric Research*, 136:95–102.

Surabhi Datta and Kirk Roberts. 2023. Weakly supervised spatial relation extraction from radiology reports. *JAMIA Open*, 6(2):ooad027.

Rodrigo del Moral-González, Helena Gómez-Adorno, and Orlando Ramos-Flores. 2025. Comparative analysis of generative LLMs for labeling entities in clinical notes. *Genomics & Informatics*, 23(1):3.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhuojun Ding, Wei Wei, and Chenghao Fan. 2025. Selecting and merging: Towards adaptable and scalable named entity recognition with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9869–9886, Vienna, Austria. Association for Computational Linguistics.

Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, and Keno K Bressem. 2024. Biomedical Large Languages Models Seem not to be Superior to Generalist Models on Unseen Medical Data. *Preprint*, arXiv:2408.13833.

David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. 2023. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, 177:105122.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Enshuo Hsu and Kirk Roberts. 2025. Leveraging large language models for knowledge-free weak supervision in clinical natural language processing. *Scientific Reports*, 15(1):8241.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, pages 1–13.

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

Yan Hu, Xu Zuo, Yujia Zhou, Xueqing Peng, Jimin Huang, Vipina K Keloth, Vincent J Zhang, Ruey-Ling Weng, Qingyu Chen, Xiaoqian Jiang, Kirk E Roberts, and Hua Xu. 2026. Information extraction from clinical notes: Are we ready to switch to large language models? *Journal of the American Medical Informatics Association*, 33(3):553–562.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and extract: Instruction tuning for on-demand information extraction. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 10030–10051, Singapore. Association for Computational Linguistics.

Veysel Kocaman, Fu-Yuan Cheng, Julio Bonis, Ganesh Raut, Prem Timsina, David Talby, and Arash Kia. 2025. Exploring named entity recognition potential and the value of tailored natural language processing pipelines for radiology, pathology, and progress notes in clinical decision support: Quantitative study. *JMIR AI*, 4(1):e59251.

Egoitz Laparra, Steven Bethard, and Timothy A Miller. 2020. Rethinking domain adaptation for machine learning over clinical language. *JAMIA Open*, 3(2):146–150.

Yinghao Li, Pranav Shetty, Lucas Liu, Chao Zhang, and Le Song. 2021. BERTifying the hidden Markov model for multi-source weakly supervised named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6178–6190, Online. Association for Computational Linguistics.

Zhiyi Li, Shengjie Zhang, Yujie Song, and Jungyeul Park. 2023. Extrinsic factors affecting the accuracy of biomedical NER. *Preprint*, arXiv:2305.18152.

Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1518–1533, Online. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P. Ma, April S. Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, Nigam H. Shah, and Jonathan H. Chen. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Meng Lu, Yuzhang Xie, Zhenyu Bi, Shuxiang Cao, and Xuan Wang. 2025. CROSSAGENTIE: Cross-type and cross-task multi-agent LLM collaboration for zero-shot information extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13953–13977, Vienna, Austria. Association for Computational Linguistics.

Marco Naguib, Xavier Tannier, and Aurélie Névéol. 2024. Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852, Miami, Florida, USA. Association for Computational Linguistics.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3):269–282.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318.

Ryan Smith, Jason A. Fries, Braden Hancock, and Stephen H. Bach. 2024. Language models in the loop: Incorporating prompting into weak supervision. *ACM / IMS J. Data Sci.*, 1(2).

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58:S20–S29. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Jinyan Su, Peilin Yu, Jieyu Zhang, and Stephen H. Bach. 2023. Leveraging large language models for structure learning in prompted weak supervision. In *Proceedings of the 2023 IEEE International Conference on Big Data (BigData)*, pages 875–884.

Xin Su, Yiyun Zhao, and Steven Bethard. 2022. A comparison of strategies for source-free domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8352–8367, Dublin, Ireland. Association for Computational Linguistics.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Han Wang, Wesley Lok Kin Yeung, Qin Xiang Ng, Angeline Tung, Joey Ai Meng Tay, Davin Ryanputra, Marcus Eng Hock Ong, Mengling Feng, and Shalini Arulanandam. 2021. A weakly-supervised named entity recognition machine learning approach for emergency medical services clinical audit. *International Journal of Environmental Research and Public Health*, 18(15):7776.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, and 16 others. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. FreeAL: Towards human-free active learning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14535, Singapore. Association for Computational Linguistics.

Yuzhen Xiao, Jiahe Song, Yongxin Xu, Ruizhe Zhang, Yiqi Xiao, Xin Lu, Runchuan Zhu, Bowen Jiang, and Junfeng Zhao. 2025. EL4NER: Ensemble learning for named entity recognition via multiple small-parameter large language models. *Preprint*, arXiv:2505.23038.

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2024. Self-improving for zero-shot named entity recognition with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 583–593, Mexico City, Mexico. Association for Computational Linguistics.

Yuki Yanagisawa, Satoshi Watabe, Sakura Yokoyama, Kyoko Sayama, Hayato Kizaki, Masami Tsuchiya, Shungo Imai, Mitsuhiro Someya, Ryoo Taniguchi, Shuntaro Yada, Eiji Aramaki, and Satoko Hori. 2025. Identifying adverse events in outpatients with prostate cancer using pharmaceutical care records in community pharmacies: Application of named entity recognition. *JMIR Cancer*, 11:e69663.

Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. 2022. A survey on programmatic weak supervision. *Preprint*, arXiv:2202.05433.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. 2021. WRENCH: A comprehensive benchmark for weak supervision. In *Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted distillation from large language models for open named entity recognition. In *Proceedings of the Twelfth International Conference on Learning Representations*, pages 1–19.

A Impact of Model Diversity

To investigate the effect of model selection on MARY, we conducted an additional study comparing our original configuration against multi-LLM setup diversified by architecture and size. Using Llama-3.3-70B as the base model, we evaluated the following configurations:

- **Architectural Diversity:** Models with different architectures: Phi-4[9] and Qwen-3-8B.[10]
- **Size Diversity:** General-purpose Llama variants: Llama-3.1-8B[11] and Llama-2-7B[12].
- **Domain Specialization:** Our original configuration adding medically-adapted (Med42-8B) and NER-specialized (UniNER-7B-type) models.

The results are summarized in Table 3. While simply adding more general-purpose models of varying architectures or sizes improves the average F1 score, the most significant gains are achieved by incorporating domain-specialized models. This suggests that MARY is most effective when the LLMs provides complementary domain knowledge.

Method	i2b2 2010	i2b2 2014	n2c2 2018	Avg.
Llama-3.3-70B	53.0	60.1	48.1	53.7
+ Qwen-3-8B, Phi-4	53.5	61.6	48.7	54.6
+ Llama-3.1-8B, Llama-2-7B	54.7	**67.5**	47.6	56.6
+ Med42-8B, UniNER-7B-type	**58.6**	65.6	**51.2**	**58.5**

Table 3: Comparison of F1 scores for different multi-LLM configurations. The best performance is highlighted in **bold**.

B Ablation Studies

We conduct an ablation to assess the contribution of each component in MARY, namely, domain-adaptive pre-training (DAPT), weighted contrastive learning, and self-training. Results are presented in Tables 4. When all components are included, MARY achieves the strongest results, demonstrating that each component independently contributes to annotation quality. Weighted contrastive learning brings consistent gains, while self-training delivers further improvements when combined with DAPT.

[9] https://huggingface.co/microsoft/phi-4
[10] https://huggingface.co/Qwen/Qwen3-8B
[11] https://huggingface.co/meta-llama/Llama-3.1-8B
[12] https://huggingface.co/meta-llama/Llama-2-7b

Method	i2b2 2010	i2b2 2014	i2b2 2018	Avg.
$\mathcal{L}_{ce}$	53.7	55.6	55.5	54.9
$\mathcal{L}_{ce} + \mathcal{L}_{ct}$	55.2	61.1	58.9	58.4
$\mathcal{L}_{ce} + \mathcal{L}_{st}$	57.2	61.8	58.9	56.8
$\mathcal{L}_{ce} + \mathcal{L}_{ct} + \mathcal{L}_{st}$	58.5	64.5	58.6	60.6
$\mathcal{L}_{ce}$+DAPT	55.0	56.7	55.7	55.8
$\mathcal{L}_{ce} + \mathcal{L}_{ct}$+DAPT	55.2	60.5	57.9	57.9
$\mathcal{L}_{ce} + \mathcal{L}_{st}$+DAPT	**58.9**	63.5	56.0	59.2
$\mathcal{L}_{ce} + \mathcal{L}_{ct} + \mathcal{L}_{st}$+DAPT	58.6	**65.7**	**59.3**	**61.2**

Table 4: Performance (micro F1) in ablation study of MARY, assessing the impact of domain-adaptive pretraining (DAPT), weighted contrastive learning ($\mathcal{L}$ct), and self-training ($\mathcal{L}$st). Cross-entropy loss ($\mathcal{L}$ce) forms the base for all settings. Each component provides a measurable gain in annotation quality, and combining all yields the best overall performance across datasets.

C Preprocessing Details

For each dataset, clinical notes in the training and test sets were split into sentences and pooled such that each sample contains at least 50 tokens. Table 5 summarizes the total numbers of original clinical notes, sentences, and pooled samples for each dataset. For zero-shot annotation evaluation, we use only the training split of each dataset. For downstream fine-tuning evaluation, the training split is used for fine-tuning, while the evaluation split is used to evaluate NER model performance. In the i2b2 2014 dataset, all entity types were merged into a single PHI category because the de-identification task aims to mask all protected health information regardless of specific type.

Dataset	i2b2 2010	i2b2 2014	n2c2 2018
Notes in Train	73	790	303
Notes in Test	353	514	202
Sentences in Train	16,388	53,064	83,321
Sentences in Test	27,757	33,967	55,322
Samples in Train	3,569	11,371	12,482
Samples in Test	6,411	7,346	8,191

Table 5: The number of clinical notes, sentences and pooled samples in train and test set for each dataset.

D More Implementation Details

D.1 Zero-shot Annotation Methods

For the vanilla zero-shot method, we generate annotations using vLLM [13] with a temperature of 0.7 and a maximum of 150 output tokens, while all other parameters use default values. For self-consistency, we repeat the vanilla zero-shot generation three times with the same parameters and retain only entities that appear in at least two runs. For self-verification, we validate the entities extracted by vanilla zero-shot by asking each model whether those entities are correct. For conflict resolution, we first identify disagreements in entity sets between Llama-3.3-70B and UniNER-7B-type, and between Llama-3.3-70B and Med-42-8B. We then use Llama-3.3-70B to verify the differing entities as in self-verification.

D.2 Few-shot Annotation Method

We follow an embedding-based retrieval method from Liu et al. (2022), where the target sample is annotated with few-shot examples that are most similar in embedding space based on cosine similarity. To retrieve similar examples, we embed both the target sample and a pool of annotated training data using a domain-adapted embedding model trained on clinical text[14] and compute cosine similarity between vectors. From a pool of 1,000 annotated training samples, we select the top 5 most similar examples for few-shot examples.

D.3 Baseline Label Models

For the HMM and CHMM, we use publicly available implementations from existing repositories.[15][16] We adopt the same hyperparameter settings as in the original implementations and do not tune them using annotated data, following Zhang et al. (2021).

D.4 MARY

The training procedure for MARY is detailed in Algorithm 1. We implement MARY using the Hugging Face Trainer.[17] For consistency and reproducibility, we utilize a fixed set of hyperparameters across all experimental datasets, as summarized in Table 6. We apply the same learning rate for both the weighted contrastive learning and self-training phases. All other hyperparameters not explicitly mentioned are kept at their default Trainer settings.

[13] https://docs.vllm.ai/en/latest/

[14] https://huggingface.co/pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb

[15] https://github.com/Yinghao-Li/CHMM-ALT

[16] https://github.com/NorskRegnesentral/skweak

[17] https://huggingface.co/docs/transformers/main_classes/trainer

Hyperparameter	Value
Learning rate	2×10^{-5}
MLM learning rate	5×10^{-5}
Weight decay	0.01
Batch size	16
Max input sequence length	350
Training epochs	5
MLM training epochs	3
Self-training epochs per iteration	1
Self-training iterations (N_{iter})	5

Table 6: Hyperparameter settings for MARY implementation.

Algorithm 1: MARY

Input : D: Unlabeled target corpus;
$\{M_1, \ldots, M_m\}$: LLMs;
L: Label model with encoder G;
$\mathcal{R}$: BIO sequence resolver;
N_{iter}: Number of self-training iterations

Output : $\hat{D}$: Aggregated annotations

`# Phase 1: Multi-LLM Annotation`

foreach $d \in D$ **do**
  Obtain LLM annotations: $\mathcal{O}_d = \{M_1(d), \ldots, M_m(d)\}$;
  Resolve to BIO sequences: $Y_d = \{\mathcal{R}(d, o) \mid o \in \mathcal{O}_d\}$;

`# Phase 2: Domain-adaptive Pre-training`

Pre-train encoder G on D using Masked Language Modeling (MLM);

`# Phase 3: Weighted Contrastive Learning`

Fine-tune L by minimizing $\mathcal{L} = \mathcal{L}_{\text{ct}} + \mathcal{L}_{\text{ce}}$ using D, $\{\mathcal{O}_d\}_{d \in D}$ and $\{Y_d\}_{d \in D}$;

`# Phase 4: Iterative Self-Training`

for $iter \leftarrow 1$ **to** N_{iter} **do**
  Predict pseudo-labels for all $d \in D$: $\tilde{\mathbf{y}} = L(d)$;
  Update L by minimizing self-training loss $\mathcal{L}_{\text{st}}$ on $\{(d, \tilde{\mathbf{y}}_d)\}_{d \in D}$;

Construct final dataset $\hat{D} = \{(d, L(d)) \mid d \in D\}$;

return $\hat{D}$;

D.5 Downstream Fine-tuning

All NER model fine-tuning is implemented using the Hugging Face Trainer. For the LLM fine-tuning, we employ Parameter-Efficient Fine-Tuning (PEFT) via LoRA (Hu et al., 2022) using the SFTTrainer from the TRL library[18].

Active Learning (AL) Baselines Our AL implementation is adapted from the source-free domain adaptation framework provided by Su et al. (2022)[19]. For both *Zero-shot + AL* and *Few-shot + AL* baselines, we first fine-tune the NER model using the parameters in Table 7 before commencing the active learning cycles. The AL process consists of 5 warm-up (WS) epochs and 5 training epochs.

LLM Fine-tuning (LLM-FT) For the LLM-FT baseline, we first fine-tune the LLM[20] using LoRA with the configuration detailed in Table 8. We then use the fine-tuned LLM to annotate the target data, which is subsequently used to train the NER model. Finally, the model is refined using manual annotations for 5 epochs following the same hyperparameter setup.

MARY+AL To jointly update MARY and the NER model, we first warm-start the NER model using automatic annotations of MARY, with hyperparameters shown in Table 7. We then proceed with active learning iterations. At each iteration, samples are selected based on the NER model's uncertainty and are used to fine-tune both MARY and the NER model. At the beginning of each iteration, we re-annotate the entire training set using the updated MARY and fine-tune the NER model with the same hyperparameters in Table 7.

Hyperparameter	Value
Learning rate	2×10^{-5}
Weight decay	0.01
Batch size	16
WS epochs (AL)	5
Training epochs (AL)	5

Table 7: General hyperparameters for NER model fine-tuning and Active Learning.

Hyperparameter	Value
Rank (r)	16
LoRA α	64
LoRA dropout	0.05
Batch size	1
Max gradient norm	0.3
Training epochs	2
Learning rate	2×10^{-5}
Optimizer	AdamW
LR scheduler	Cosine
Warmup ratio	0.05
Max sequence length	3000

Table 8: Hyperparameters for LLM fine-tuning using LoRA and SFT.

[18] https://huggingface.co/docs/trl/sft_trainer

[19] https://github.com/xinsu626/SourceFreeDomainAdaptation

[20] https://huggingface.co/meta-llama/Llama-3.1-8B

E Discussion on Performance of Baseline Label Models

E.1 Precision–Recall Trade-off of Majority Voting

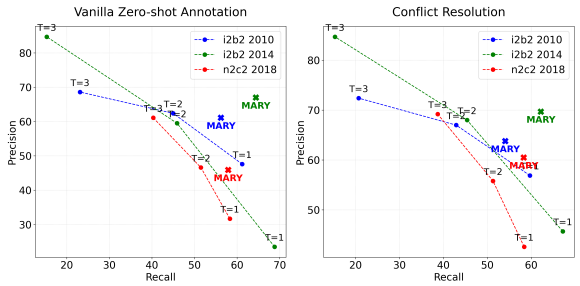

Figure 6: Precision–recall trade-off of MV baselines with different thresholds ($T = 1, 2, 3$) and MARY. The left shows vanilla zero-shot annotation, while the right shows results with conflict resolution. The increase in recall from $T = 2$ to $T = 1$ indicates that incorporating minority-extracted entities is essential for comprehensive annotation. MARY achieves recall higher than MV ($T = 2$) while maintaining comparable precision.

Fig. 6 illustrates the precision–recall trade-off of majority voting (MV) under different agreement thresholds ($T = 1, 2, 3$). Across both vanilla zero-shot annotation and conflict resolution, reducing the threshold from $T = 2$ to $T = 1$ leads to a substantial increase in recall across all datasets, indicating that entities extracted by only a minority of models account for a significant portion of true positives. However, this recall gain comes at the cost of reduced precision.

In contrast, MARY achieves recall higher than MV ($T = 2$) while maintaining precision at a comparable level, resulting in consistently higher F1 scores than MV baselines. This demonstrates that selectively incorporating minority-extracted entities based on contextual consistency is more effective than uniformly lowering the agreement threshold.

E.2 Underperformance of HMM and CHMM

In our evaluation, HMM and CHMM only occasionally improve over the best-performing single-model annotation (Llama-3.3-70B-Instruct), despite their effectiveness in prior work using conventional labeling functions such as dictionaries. We hypothesize that this limitation arises from the inherently noisy nature of LLM annotations.

Fig. 7 illustrates the relationship between (i) the average precision and recall of LLM annotations and (ii) the resulting F1 scores of the label models (HMM and CHMM). Improvements in precision achieved via zero-shot refinement (i.e., conflict resolution) consistently lead to proportional gains in F1 for both label models. In contrast, improvements in recall exhibit no correlation with F1 performance. This pattern indicates that HMM and CHMM are substantially more sensitive to annotation noise than to annotation coverage.

Unlike traditional rule-based labeling functions, which are typically high-precision but low-coverage, LLM annotation offers greater flexibility at the expense of increased noise. As a result, label models that assume relatively clean supervision, such as HMM and CHMM, underperform under low-precision annotations, highlighting a key limitation when applying these models to noisy LLM annotations.

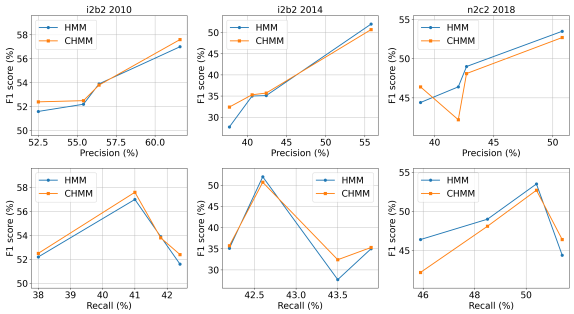

Figure 7: Relationship between annotation quality and label model performance. Precision–F1 (top row) and recall–F1 (bottom row) plots for HMM and CHMM across datasets. Each point corresponds to a different zero-shot annotation method, ordered by increasing precision or recall. Improvements in annotation precision lead to consistent and near-monotonic gains in F1 for both HMM and CHMM. In contrast, changes in annotation recall show no clear relationship with F1 score.

F Majority Voting with All Models vs. MARY

To examine whether larger ensembles with simple aggregation can better capture minority entities, we evaluate majority voting (MV) using all seven LLMs. The results are presented in Table 9. Overall, MV exhibits inconsistent performance across thresholds and datasets, and the performance does not reliably improve with increasing ensemble size. Even when using all seven models, MV fails to consistently outperform smaller ensembles. In contrast, MARY achieves the highest average F1 score, demonstrating more stable and robust performance. These findings indicate that simply increasing the number of models is insufficient for improving ex-

traction quality. Instead, more sophisticated aggregation methods, such as MARY, are necessary to effectively integrate diverse model outputs.

G Inference Cost and Practical Trade-offs

We analyze the computational cost of each model in terms of token usage and inference time. Table 10 summarizes average prompt tokens, generated tokens, per-batch inference time, and estimated total inference time for annotating the full training split. Using Llama-3.3-70B as the primary annotator incurs the highest computational cost. However, augmenting it with smaller models (UniNER-7B-type and Med42-8B) increases total inference time by only 8–25%, while yielding a substantial improvement in extraction performance (approximately +5.9 F1 on average). This result suggests that combining a large general-purpose model with smaller, specialized models provides a favorable balance between performance and computational efficiency.

H Prompt Templates

Our prompt templates are distilled directly from the official annotation guidelines of each dataset. Specifically, we provide the LLMs with summarized descriptions of each clinical entity category to mitigate ambiguity. Representative prompt templates for each dataset are illustrated in Figs. 8, 9, and 10.

LLMs Used	Aggregation	i2b2 2010	i2b2 2014	n2c2 2018	Avg.
Med42-8B, UniNER-7B-type, Llama-70B	MV (T=1)	53.6	35.1	41.1	43.2
Med42-8B, UniNER-7B-type, Llama-70B	MV (T=2)	52.3	51.8	48.9	51.0
Med42-8B, UniNER-7B-type, Llama-70B	MV (T=3)	34.6	25.9	48.6	36.4
Qwen, Phi-4, Llama-70B	MV (T=1)	50.5	32.6	44.1	42.4
Qwen, Phi-4, Llama-70B	MV (T=2)	52.1	55.1	49.5	52.2
Qwen, Phi-4, Llama-70B	MV (T=3)	46.1	37.1	51.2	44.8
Llama-8B, Llama-7B, Llama-70B	MV (T=1)	45.3	24.2	37.0	35.5
Llama-8B, Llama-7B, Llama-70B	MV (T=2)	51.0	46.9	46.7	48.2
Llama-8B, Llama-7B, Llama-70B	MV (T=3)	46.7	11.7	43.3	33.9
All 7 LLMs	MV (T=1)	44.6	19.7	33.9	32.7
All 7 LLMs	MV (T=2)	53.5	35.8	44.3	44.5
All 7 LLMs	MV (T=3)	56.0	48.4	48.6	51.0
All 7 LLMs	MV (T=4)	55.5	50.0	51.1	52.2
All 7 LLMs	MV (T=5)	52.0	39.4	**52.5**	48.0
All 7 LLMs	MV (T=6)	47.1	18.9	52.2	39.4
All 7 LLMs	MV (T=7)	39.4	2.7	47.3	29.8
Med42-8B, UniNER-7B-type, Llama-70B	MARY	**58.6**	**65.6**	51.2	**58.5**

Table 9: Performance comparison across LLM combinations and aggregation methods

Dataset	Model	Input Tokens	Gen Tokens	Per Batch (sec)	Total (hr)
i2b2 2010	Llama-3.3-70B	1557.37	158.20	17.97	4.46
	UniNER-7B-type	502.64	112.20	0.499	0.12
	Med42-8B	1557.37	160.36	0.941	0.23
i2b2 2014	Llama-3.3-70B	1485.28	52.92	5.91	4.66
	UniNER-7B-type	573.05	84.92	0.405	0.32
	Med42-8B	1485.28	83.76	0.510	0.40
n2c2 2018	Llama-3.3-70B	1717.91	196.00	21.18	18.36
	UniNER-7B-type	594.70	166.36	0.685	0.59
	Med42-8B	1717.91	436.44	4.817	4.18

Table 10: Inference statistics across datasets and models. The results are averaged over 500 samples.

<|system|> **You are a medical expert.** Your task is to identify and mark **Medical Problems** in clinical text.

Guidelines:

Medical Problems
Medical Problems are diseases, symptoms, signs, injuries, infections, or abnormal findings. They must be things wrong with the patient that can be treated or diagnosed. Do not include normal states, measurements, or habits (e.g., smoking, alcohol).

Tests
Tests are lab procedures, diagnostic procedures, examinations, or measures. They are used to discover or rule out problems. Do not include test results or values.

Treatments
Treatments are procedures, interventions, medications, or devices used to fix a problem. Include drugs (brand or generic), procedures, surgeries, devices, or therapy types. Do not include verbs or people or locations giving treatment.

Your final output should be a list containing only the medical problem entities present in the document. Please follow the format below exactly.

`Medical Problems entities: ["entity1", "entity2", ...]`

Extracted entities must be complete noun phrases (NPs) or adjective phrases (APs). A noun phrase is a group of words that functions as a noun in a sentence, typically consisting of a noun and its modifiers, such as adjectives, determiners, or other descriptors. For example, "the tall building near the park" functions as a single noun.

Do not include any explanations, especially inside the list.
<|end|>
<|user|>text<|end|>
<|assistant|>

Figure 8: Zero-shot prompt template for i2b2 2010 dataset. For i2b2 2010 dataset, descriptions for all entity types are included since we observed that LLM struggle with distinguising test and treatment entity due to ambiguous entity type boundry. The output is formatted as list of entities.

<|system|>
You are a medical / clinical NLP expert.
Your task is to identify and mark **PHI (Protected Health Information)** entities in clinical text. **PHI** includes any information that could directly or indirectly identify a patient, such as names, dates, locations, contact information, IDs, ages, and any other identifiers.

Annotation Guidelines:

- Mark only the sensitive text that would be removed or replaced during de-identification.
- Include all ages, all parts of dates (day, month, year), days of week, seasons, room numbers, floors, suites, and any unique identifiers.
- Include names of patients, doctors, staff, usernames, job titles, and any other identifying labels.
- Include addresses, hospital names, departments, cities, states, countries, ZIP codes, and other location references.
- Include phone numbers, fax numbers, emails, URLs, IP addresses, social security numbers, medical record numbers, health plan IDs, account numbers, license numbers, vehicle IDs, device IDs, biometric IDs, or any other ID numbers.

Your final output should be a list containing only the PHI entities present in the document. Please follow the format below exactly.

`PHI entities: ["entity1", "entity2", ...]`

Do not include any explanations, especially inside the list.
<|end|>
<|user|>text<|end|>
<|assistant|>

Figure 9: Zero-shot prompt template for PHI (Protected Health Information) identification. The prompt provides comprehensive guidelines on sensitive identifiers including dates, names, locations, and unique IDs.

<|system|>
You are a medical expert.
Your task is to identify **ADE** (Adverse Drug Event) entities in clinical text.

Guidelines:

ADE
This entity type should be selected to specify an **Adverse Drug Event (ADE)**. This includes any injury resulting from medical intervention related to a drug, such as side effects, hypersensitivity, or toxicities.

Examples:
1. Patient is experiencing **muscle pain**, secondary to statin therapy.
`ADE entities: ["muscle pain"]`

2. The patient suffers from steroid-induced **hyperglycemia**.
`ADE entities: ["hyperglycemia"]`

3. Patient prescribed 1–2 325 mg/10 mg Norco pills for pain.
`ADE entities: []`

Your final output should be a list containing only the ADE entities present in the document. Please follow the format below exactly.

`ADE entities: ["entity1", "entity2", ...]`

If no ADE entities are found, respond with an empty list: `[]`.
Do not include any explanations, especially inside the list.
<|end|>
<|user|>text<|end|>
<|assistant|>

Figure 10: Zero-shot prompt template for Adverse Drug Event (ADE) extraction. The prompt provides negative examples (prescriptions without adverse events) derived directly from the annotation guideline to help the model distinguish between drug mentions and actual drug-induced injuries.