Evaluating Clinical NER in Few-shot Settings

Tianxin Wang **Yating Xu** **Yiyao Zhang** **Chenxin Gu**
Center for Data Science
New York University
{tw3090, yx3710, yz12490, cg3423}@nyu.edu

1 Motivation & Prior Work

Healthcare worldwide faces persistent shortages, and much clinical knowledge remains in unstructured text. Extracting entities such as diseases and symptoms through Named Entity Recognition (NER) is essential for applications like decision support and patient profiling. However, annotation is costly, privacy sensitive, and time-consuming, making clinical NER inherently a small-data problem.

Early rule-based and supervised systems struggled with the ambiguity and jargon of clinical text, while annotated corpora remain scarce [1, 2, 3]. Transfer learning with models such as BioBERT and ClinicalBERT has since become standard, and prompting with large LLMs has been explored, although performance is unstable and often inferior to masked LMs in realistic few-shot settings [4, 5, 6]. Building on this line of work, we will systematically evaluate **metric learning, meta-learning, transfer learning, and prompting** in the multilingual E3C corpus to provide a more complete picture of clinical NER with a few shots. In particular, we aim to directly compare the performance of these paradigms under consistent few-shot settings to identify their relative strengths and weaknesses.

2 Datasets

Primary dataset: E3C (English subset). We use the English portion of the European Clinical Case Corpus (E3C), which consists of approximately 800 clinical case reports. The dataset is annotated for multiple types of clinical entity, including *diseases, treatments, medications, symptoms*, and *tests*. The data is distributed in **CoNLL-style format** (token–label pairs). In this project, we restricted our study to the English subset in order to focus on the clinical NER task under few-shot conditions.

Backup dataset: NCBI Disease. As a supplementary resource, we include NCBI Disease corpus, which contains roughly 7,000 PubMed sentences annotated for mentions of *diseases*. The corpus is also distributed in **CoNLL-style format**. Compared to E3C, it is larger in size and originates from biomedical literature rather than real clinical narratives, and therefore serves as a complementary benchmark when the small size of E3C limits experimental robustness.

3 Methods

Metric learning. We include **Prototypical Networks** as a representative metric learning approach. The key idea is to embed tokens into a latent space and compute class prototypes by averaging support examples. During inference, new spans are classified based on the nearest prototype in this embedding space. Metric learning is especially suitable in the few-shot regime, as it does not require full parameter updates and instead relies on distance comparisons. In the context of clinical NER, this paradigm should provide a strong baseline for recognizing new entity types from very limited examples.

Meta-learning. We adopt **MAML** (Model-Agnostic Meta-Learning) to represent meta-learning strategies. Meta-learning trains a model across many few-shot "episodes" so that it learns how to adapt quickly to unseen tasks with only a handful of gradient updates. This is particularly relevant for clinical NER, where new entity categories (e.g., a rare disease or treatment) may appear without sufficient annotated data. By simulating this adaptation process, meta-learning should offer a principled way to evaluate rapid generalization in the small-data setting.

Transfer learning. We evaluate transfer learning via **BioBERT** , pretrained language models widely used in biomedical NLP. Both full fine-tuning and **parameter-efficient tuning (LoRA, prefix-tuning)** are compared. Transfer learning has become the de facto approach for low-resource NER tasks, as pretrained models bring domain knowledge from large corpora such as PubMed or MIMIC. Our study will assess how these methods behave in strict few-shot conditions and whether lightweight PEFT methods can provide more stable results than full fine-tuning.

Prompting. We include prompting with **large language models** such as GPT-4 and Flan-T5. Rather than training new parameters, prompting provides the model with task instructions and a few labeled examples in-context. We will experiment with both *minimal templates* (short examples only) and *enriched templates* (examples plus label definitions). By benchmarking prompting against other paradigms on clinical data, we hope it has practical value in real few-shot clinical NER scenarios.

4 Goal & Evaluation

The aim of this project is to evaluate the feasibility of performing few-shot clinical Named Entity Recognition (NER) in a small-data environment. Using the English subset of the E3C corpus, we seek to determine how well NER can be achieved when only a limited number of annotated examples are available, reflecting real-world constraints in healthcare text analysis.

The evaluation will focus on the following dimensions:

- **Entity-level F1-score:** primary accuracy metric balancing precision and recall.
- **Data efficiency:** rate of performance improvement with additional examples.
- **Computational cost:** resources required for training and inference (time, memory).
- **Robustness:** stability across runs and sensitivity to random initialization.

In addition, we will compare the results across the four methods to highlight their relative strengths and weaknesses under few-shot conditions.

5 Allocation of Work & Timeline

Allocation of Work.

For focus and efficiency, each team member will be primarily responsible for one methodological paradigm. Tianxin Wang will work on metric learning, Yating Xu on meta-learning, Yiyao Zhang on transfer learning, and Chenxin Gu on prompting.

Timeline.

Step 1: Data Preparation (Week 1: Sep 24 – Sep 30)

- Preprocess the E3C English subset and convert it into a standard NER format (IOB/IOB2).

Step 2: Baseline Implementation (Week 2–4: Oct 1 – Oct 14)

- Run baseline models from different paradigms with minimal tuning on preprocessed data.

Step 3: Advanced Experiments & Analysis (Week 5–7: Oct 15 – Nov 7)

- Extend experiments by applying parameter-efficient fine-tuning, designing enriched prompts, and analyzing data efficiency, robustness, and scalability. Consolidated results and figures for report and presentation.

Step 4: Presentation Preparation (Week 8: Nov 8 – Nov 14)

- Finalize a clear and concise 10-minute presentation.

Step 5: Final Presentation (Nov 18 – Dec 9)

- Deliver in-class presentation.

References

[1] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. Design and implementation of a multipurpose clinical natural language processing system. *Journal of the American Medical Informatics Association*, 16(5):799–805, 2009.

[2] Robert Leaman and Zhiyong Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37, 2015.

[3] Yuan Luo, Xia Sun, Anna Rumshisky, Özlem Uzuner, and Peter Szolovits. An overview of the clinical named entity recognition methods. *Methods*, 166:3–13, 2020.

[4] Zhengxuan Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

[5] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Rethinking the role of demonstrations: What makes in-context learning work? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[6] Mohamed Zaghir et al. Few-shot clinical entity recognition in english, french and spanish: masked language models outperform generative model prompting. *arXiv preprint arXiv:2402.12801*, 2024.