WiseNER: Meta-reasoning for Large Language Models in Biomedical Named Entity Recognition

Hang Zhang[1,2#], Wendong Ge[1], Ruheng Wang[1,3#], Gregory Bryman[1]*, Antong Chen[1]
[1]Merck & Co., Inc., Rahway, NJ, USA
[2]University of Pittsburgh, Pittsburgh, PA, USA
[3]UT Southwestern Medical Center, Dallas, TX, USA
[#] contributed to this work as interns when working in Merck.
* Corresponding author (Email: gregory.bryman@merck.com)

1. Introduction

The exponential growth of unstructured biomedical text, encompassing scientific literature, clinical trial records, electronic health records (EHRs), and patent filings, has created an unprecedented wealth of information. However, the vast majority of this data remains inaccessible to automated systems, creating a significant bottleneck in biomedical research and healthcare. Biomedical Named Entity Recognition (BiomedNER) serves as a cornerstone technology to address this challenge. By transforming unstructured text into structured, machine-readable data, BiomedNER enables a wide array of downstream applications, including advanced literature reviews, automated knowledge base construction, drug discovery, target mining, non-clinical drug safety analysis, and precision medicine. [1]

Despite its critical importance, the development of high-performing BiomedNER systems is perennially constrained by the availability of high-quality labeled data. The process of annotating biomedical text is notoriously expensive, labor-intensive, and requires significant domain expertise. For out-of-domain entities without sufficient labeled data, the current BiomedNER systems do not work well. [2]

Thus, in this work, we proposed WiseNER, a novel two-stage framework that addresses the zero-shot or one-shot generalization challenge in BiomedNER through Large Language Models (LLMs), Meta-Reasoning Distillation and Reinforcement Learning.

2. Methods

Our approach leveraged **Meta-Reasoning Patterns** to transform entity recognition into a structured reasoning problem, followed by **Reasoning Trace Distillation** and **Reinforced Fine-tuning** to solidify and generalize the learned reasoning capabilities.

Meta Reasoning Patterns: We reformulated the traditional NER task as a reasoning problem by introducing a meta-reasoning pattern that explicitly decomposes entity identification into systematic cognitive steps. This approach moved beyond pattern matching to engage the model in deliberative reasoning about entity boundaries and types. Our meta-reasoning pattern employed a dual-strategy framework consisting of RECALL and PRECISION phases. The RECALL strategy performed an exhaustive scan of the input text to identify all potential entity candidates, including partial matches,

compound terms, and coordinated structures. The PRECISION strategy then validated each candidate against the entity type definition, determining exact boundaries and confirming contextual appropriateness

Reasoning Trace Distillation: The reasoning trace distillation stage formed the foundation of our framework by teaching a lightweight model to internalize the systematic reasoning process for entity identification. We leveraged a more capable teacher model to generate detailed reasoning traces that explain the step-by-step logic behind correct entity annotations.

Reinforced Fine-tuning: Following the reasoning trace distillation, we employed reinforced fine-tuning to further refine and generalize the model's reasoning capabilities. This stage used reinforcement learning to encourage the model to generate high-quality reasoning traces that lead to correct entity identifications, even in challenging or ambiguous cases.

3. Results

We selected Llama-3.1-8B [3], a small open-source LLM, to build our WiseNER framework, and compared it with some leading closed-sourced big LLMs such as GPT-4o and o3. In the Reasoning Trace Distillation stage and Reinforced Fine-tuning stage, we used drug name and disease from BC5CDR dataset [4] as in-domain entity types to post-train the LLM; In the inference stage, we used Organism as out-of-domain entity type from BioRED dataset [5] for benchmarking.

Firstly, for zero-shot prompting, original Llama-3.1-8B achieved 0.263 F1 score, while GPT-4o and o3 achieved 0.470 and 0.498 F1 scores. After post-training with our WiseNER framework, Llama-3.1-8B achieved 0.483 F1 score, which beats GPT-4o and is close to o3. Additionally, for one-shot prompting, original Llama-3.1-8B achieved 0.515 F1 score, while GPT-4o and o3 achieved 0.719 and 0.653 F1 scores respectively. After post-training with our WiseNER framework, Llama-3.1-8B achieved 0.763 F1 score, which beats both GPT-4o and o3.

4. Conclusions

This work presented WiseNER, a novel framework for enhancing zero-shot or one-shot generalization in biomedical Named Entity Recognition through meta-reasoning and reinforced fine-tuning. Our approach gave an option to help address the fundamental challenge of data scarcity in specialized biomedical domains by teaching models to reason systematically about entity identification rather than relying on memorized **patterns**. Our empirical results demonstrated that this framework enabled smaller models to achieve performance competitive with or superior to much larger baseline models, while showing substantial improvements in zero-shot or one-shot generalization to novel entity types.

Acknowledgements and Disclosures

The authors declare that they have no financial relationships or other potential conflicts of interest relevant to this research.

Reference

[1] Song B, Li F, Liu Y, Zeng X. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. Briefings in Bioinformatics. 2021 Nov;22(6):bbab282.
[2] Chen, Qingyu, et al. "Benchmarking large language models for biomedical natural language processing applications and recommendations." Nature communications 16.1 (2025): 3280.
[3] Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Vaughan A, Yang A. The llama 3 herd of models. arXiv preprint arXiv:2407.21783. 2024 Jul 31.
[4] Li J, Sun Y, Johnson RJ, Sciaky D, Wei CH, Leaman R, Davis AP, Mattingly CJ, Wiegers TC, Lu Z. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database. 2016 Jan 1;2016.
[5] Luo L, Lai PT, Wei CH, Arighi CN, Lu Z. BioRED: a rich biomedical relation extraction dataset. Briefings in Bioinformatics. 2022 Sep;23(5):bbac28