World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:18, No:10, 2024

Open Science Index, Cognitive and Language Sciences Vol:18, No:10, 2024 publications.waset.org/abstracts/189305.pdf

# Little Retrieval Augmented Generation for Named Entity Recognition: Toward Lightweight, Generative, Named Entity Recognition Through Prompt Engineering, and Multi-Level Retrieval Augmented Generation

**Authors :** Sean W. T. Bayly, Daniel Glover, Don Horrell, Simon Horrocks, Barnes Callum, Stuart Gibson, Mac Misuira

**Abstract :** We assess suitability of recent, ~7B parameter, instruction-tuned Language Models Mistral-v0.3, Llama-3, and Phi-3, for Generative Named Entity Recognition (GNER). Our proposed Multi-Level Information Retrieval method achieves notable improvements over finetuned entity-level and sentence-level methods. We consider recent developments at the cross roads of prompt engineering and Retrieval Augmented Generation (RAG), such as EmotionPrompt. We conclude that language models directed toward this task are highly capable when distinguishing between positive classes (precision). However, smaller models seem to struggle to find all entities (recall). Poorly defined classes such as ”Miscellaneous” exhibit substantial declines in performance, likely due to the ambiguity it introduces to the prompt. This is partially resolved through a self verification method using engineered prompts containing knowledge of the stricter class definitions, particularly in areas where their boundaries are in danger of overlapping, such as the conflation between the location ”Britain” and the nationality ”British”. Finally, we explore correlations between model performance on the GNER task with performance on relevant academic benchmarks.

**Keywords :** generative named entity recognition, information retrieval, lightweight artificial intelligence, prompt engineering, personal information identification, retrieval augmented generation, self verification
**Conference Title :** ICNLCLP 2024 : International Conference on Natural Language Computing and Language Processing
**Conference Location :** Barcelona, Spain
**Conference Dates :** October 24-25, 2024