Few-Shot Prompting with Masked Language Models vs. Large Autoregressive Models for Low-Resource Clinical Named Entity Recognition

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does few-shot prompting with lightweight masked language models compare to large autoregressive models on low-resource clinical named entity recognition benchmarks. 13 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting. Research question: How does few-shot prompting with lightweight masked language models compare to large autoregressive models on low-resource clinical named entity recognition benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

11 papers retrieved. 13 claims extracted; 2 independently verified. Quality review score: 4.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The study evaluates 14 NER tasks spanning 6 general-domain datasets and 8 clinical datasets.	×	0.11
The study focuses on clinical entity recognition in English, French, and Spanish.	✓	0.21
The study compares 13 generative LLMs against 16 fine-tuned Masked Language Models (MLMs).	✓	0.16
Prompt optimization was performed only on few annotated instances through cross-validation to ensure a true few-shot set	×	0.03
On the English WikiNER dataset, the XLM-R-large model achieved an F1 score of 0.826 with p=1.	×	0.05
On the English WikiNER dataset, the Mistral-7B model achieved an F1 score of 0.646 with p=1.	×	0.04
On the French E3C dataset, the XLM-R-large model achieved an F1 score of 0.462 with p=1.	×	0.04
On the French E3C dataset, the Mistral-7B model achieved an F1 score of 0.291 with p=1.	×	0.04
Llama-2-70B achieved an F1 score of 0.728 on the English WikiNER dataset.	×	0.04
Phi-3-medium-instruct achieved an F1 score of 0.464 on the English WikiNER dataset.	×	0.04
Vigogne-13B achieved an F1 score of 0.593 on the English WikiNER dataset.	×	0.04
The study identifies that most recent studies employing prompt engineering in medical applications lack a non-prompt-rel	×	0.06
The code for the study is available at github.com/marconaguib/autoregressive_ner.	×	0.02

References

- http://arxiv.org/abs/2210.12770v4
- http://arxiv.org/abs/2302.04725v1
- http://arxiv.org/abs/2402.12801v2