Intelligent Health Systems – From Technology to Data and Knowledge
E. Andrikopoulou et al. (Eds.)

doi:10.3233/SHTI250492

Delirium Identification from Nursing Reports Using Large Language Models

Lisa GRAF[a,b,1], Alexander RITZI[c,d], and Lili M. SCHOELER[c,e]
[a] *Neurorobotics Lab, Department of Computer Science - University of Freiburg, Germany*
[b] *Department of Neurology, Medical Center - University of Freiburg, Germany*
[c] *Center of Implementing Nursing Care Innovations Freiburg, Medical Center – University of Freiburg, Germany*
[d] *Centre for Geriatric Medicine and Gerontology (ZGGF), Medical Center – University of Freiburg, Germany*
[e] *Department of Nursing Science, University of Turku, Finland*
ORCiD ID: Lisa Graf 0009-0008-3623-1480, Alexander Ritzi 0009-0003-6502-4118, Lili M. Schoeler 0000-0002-6850-3179

Abstract. This study investigates large language models for delirium detection from nursing reports, comparing keyword matching, prompting, and finetuning. Using a manually labelled dataset from the University Hospital Freiburg, Germany, we tested Llama3 and Phi3 models. Both prompting and finetuning were effective, with finetuning Phi3 (3.8B) achieving the highest accuracy (90.24%) and AUROC (96.07%), significantly outperforming other methods.

Keywords. Electronic Health Records, Large Language Models, Delirium

1. Introduction

Delirium is a serious, often under-reported condition. Prediction models face challenges from poor-quality labels: ICD codes underestimate incidence and lack timestamps, while assessments capture only specific moments, missing interim cases. Nursing reports offer continuous shift-by-shift labels, but the classification requires more than word matching due to variability in how staff recognize and describe delirium. This study investigates using large language models to classify nursing reports.

2. Methods

Two nursing scientists manually labelled the German reports based on their practical experience with delirious patients (Cohens Kappa of 0.82 based on 277 reports). The dataset comprised 851 training samples and 182 each for validation and testing (54.64% positive cases). We compared zero-shot, one-shot, and few-shot prompting strategies to keyword matching with manually selected keywords, and finetuning with qLoRA [3], replacing the final layer with a classification head. For few-shot prompting, the model

[1] Corresponding Author: Lisa Graf; E-mail: lgraf@cs.uni-freiburg.de.

was given two positive and one negative example. The prompting templates were too long to include in this paper. Experiments used Llama3 [1] and Phi3 [2] models.

3. Results

The prompting and finetuning performance are summarized in Table 1. As a baseline, keyword matching achieved an accuracy of 80.27%, but with a low specificity of 63.75%.

Table 1. Performance metrics for prompting and finetuning across different models (Acc.: accuracy, Sens.: Sensitivity, Spec.: Specificity), qLoRA setup: r=256, alpha=256, LoRA dropout: 0.05, and DoRA [4] enabled. Finetuning metrics were averaged over three runs.

Prompting	**Zero Shot**	**One Shot**	**Few Shot**
	Acc./Sens./Spec. [%]	**Acc./Sens./Spec. [%]**	**Acc./Sens./Spec. [%]**
Llama3 (8B)	80.74/71.92/90.48	80.86/75.74/87.28	77.30/88.89/69.23
Llama3 (70B)	87.11/86.92/87.31	85.89/76.76/95.27	87.73/82.81/92.79
Phi3 (3.8B)	16.32/24.27/98.18	34.48/81.40/79.10	75.71/86.65/66.84
Finetuning	**Accuracy [%]**	**AUROC [%]**	**Precision/Recall [%]**
Llama3 (8B)	87.07	95.05	87.24/87.07
Llama3German (8B)	88.16	95.51	88.37/88.16
Phi3 (3.8B)	90.24	96.07	90.22/90.07

4. Discussion

Among the prompting methods, Llama3 (70B) achieved the best accuracy (87.73%) with few-shot prompting. Phi3 initially performed poorly in zero-shot prompting (16.32% accuracy) but improved to 75.71% with few-shot prompting. Finetuning significantly increased accuracy for all models, with Phi3 achieving the highest accuracy (90.24%) and AUROC (96.07%), while Llama3 (8B) reached 87.07% accuracy and Llama3 German (8B) reached 88.16% accuracy. Overall, finetuning consistently outperformed prompting when comparing models of the same size. Prompting showed substantial variability depending on the examples provided, highlighting finetuning's advantage in eliminating prompt optimization overhead.

5. Conclusions

Our experiments indicate that while prompting is useful, finetuning delivers more reliable performance for delirium detection. Future work will extend experiments and incorporate explainability methods.

References

[1] Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Ganapathy R. The Llama 3 Herd of Models. arXiv preprint; 2024.
[2] Abdin M, Jacobs SA, Awan AA, Aneja J, Awadallah A, Awadalla H, Zhou X. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint; 2024
[3] Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. Qlora: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems. 2024.
[4] Liu SY, Wang CY, Yin H, Molchanov P, Wang YCF, Cheng KT, Chen MH. Dora: Weight-decomposed low-rank adaptation. arXiv preprint; 2024