Evaluating Machine Learning Approaches for Multi-Label Classification of Unstructured Electronic Health Records with a Generative Large Language Model

Dinithi Vithanage, Chao Deng, Lei Wang, Mengyang Yin, Mohammad Alkhalaf, Zhenyua Zhang, Yunshu Zhu, Alan Christy Soewargo, Ping Yu.

University of Wollongong, Australia.

1. Research Problem

❖**Early Stage:**
- **Generative AI-based LLMs are still in the early stages of being applied to extract clinical insights from free-text electronic health records (EHR).**

❖**Potential vs. Limitations:**
- **LLMs have shown promise in answering clinical questions and extracting data from public health datasets, but their application in real-world clinical tasks remains limited.**

❖**Safety Concerns:**
- **The ability of LLMs to meet stringent healthcare safety standards is uncertain, with risks of generating disinformation, bias, or hallucinations.**

❖**Prompting Strategies:**
- **The optimal prompting strategies for healthcare information extraction (zero-shot vs. few-shot) are still unclear.**

2. Research Aim

❖**The experimental research aims to test the effect of the zero-shot and few-shot learning prompting strategies, with and without retrieval augmented generation (RAG) and parameter efficient fine-tuning (PEFT) LLMs, on the multi-label classification of the EHR data set.**

3. Research Methodology

1. Obtain ethics approval

2. Select the Llama 3-8B parameter model

3. Collect data from residential aged care facilities

4. Select clinical tasks for multi-label classification:
- **agitation in dementia**
- **depression in dementia**
- **frailty index**
- **malnutrition risk factors**

5. Execute machine learning methods:
- **zero-shot and few-shot prompt-based learning**
- **PEFT**
- **RAG**

6. Evaluate model performance using accuracy, precision, recall, and F1 score.

7. Conduct statistical analysis

4. Results

- **The same level of performance with the same prompting template, either zero-shot or few-shot learning across the four clinical tasks.**
- **Few-shot learning outperforms zero-shot learning without PEFT.**
- **Fine-tuning significantly enhanced the effectiveness of both zero-shot and few-shot learning.**
- **The performance of zero-shot learning reached the same level as few-shot learning after PEFT.**
- **The analysis underscores that LLMs with PEFT for specific clinical tasks maintain their performance across diverse clinical tasks.**
- **The RAG with few-shot learning outperforms RAG with zero-shot learning, while there is no significant difference between RAG with few-shot and PEFT with zero-shot learning.**

5. Conclusion

- **RAG with few-shot learning and PEFT with zero-shot or few-shot learning plays a crucial role in optimizing LLM performance.**
- **These insights emphasise the adaptability and effectiveness of RAG and PEFT within the LLMs for various clinical tasks.**