Detecting Adverse Events and Geriatric Syndromes in Clinical Text: A Preliminary Study on MIMIC-IV Using Large Language Models

Fahrurrozi Rahman, Aryo Pradipta Gema, Bruce Guthrie, Beatrice Alex
University of Edinburgh

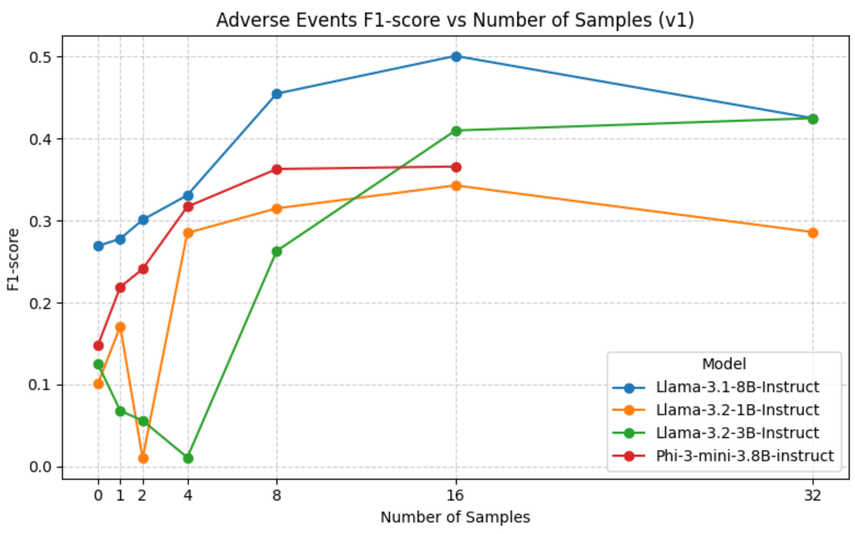
INTRODUCTION

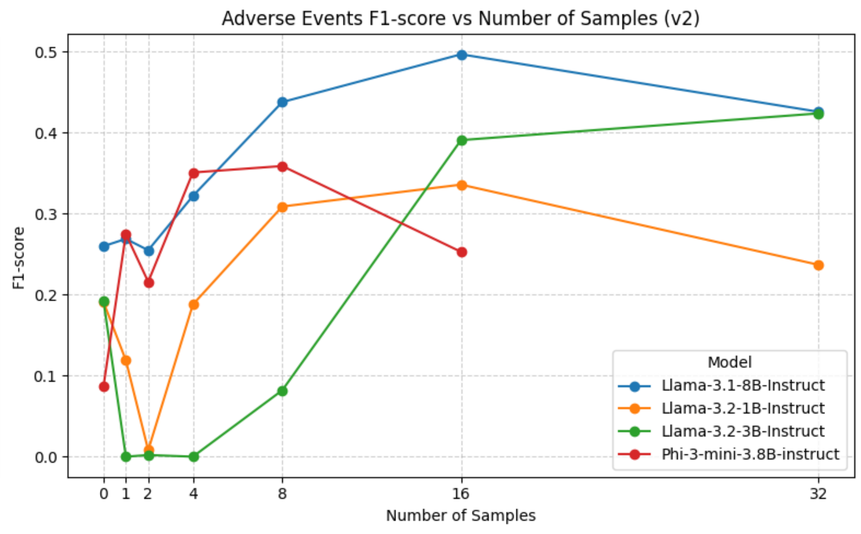
- Adverse events (AEs) are unintended injuries or complications that arise from medical care or occur during hospitalisation.
- Geriatric syndromes (GS), such as falls, delirium, frailty, etc., are highly prevalent in older populations and strongly associated with poor health outcomes
- Detecting AEs and GS automatically from clinical narratives can improve population-level surveillance, support research on multi-morbidity, and enable more responsive clinical decision support.
- Progress in this area remains limited due to the scarcity of annotated data and the heterogeneity of clinical language across healthcare systems.
- Unstructured data in electronic health records (EHRs) contain rich information that remains underutilised.
- Large Language Models (LLMs) can be leveraged for document classification tasks through in-context learning (ICL) without additional training.

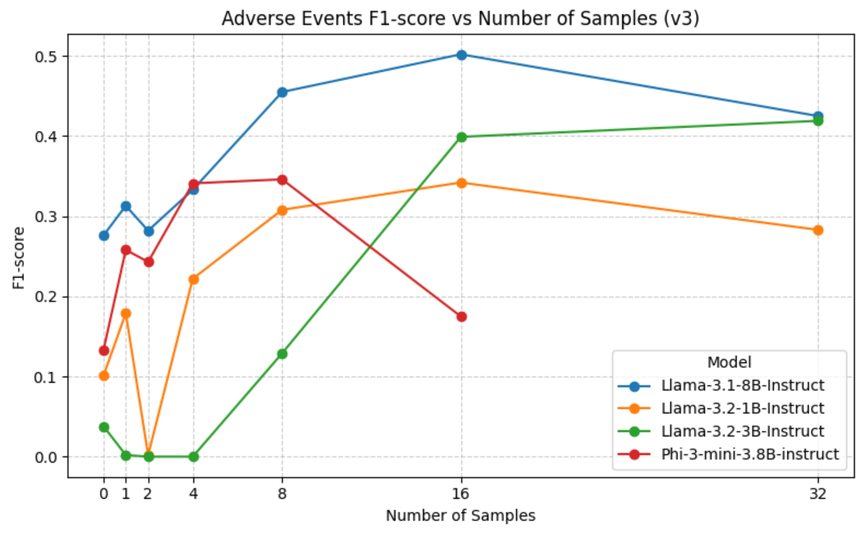
METHODS

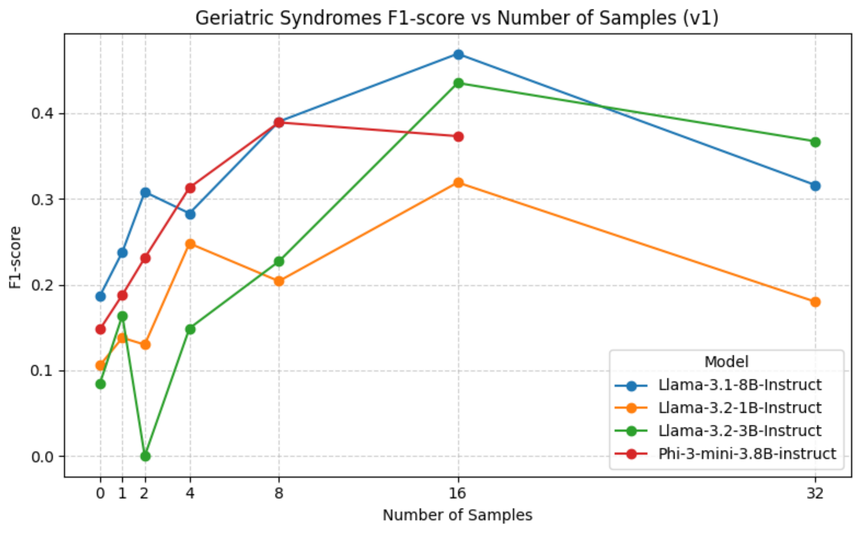
- We evaluated ICL strategies for multi-label document classification covering 14 AEs, 12 GS, and their negations in annotated MIMIC-IV discharge summaries [1].
- We selected Llama and Phi model families for their openness, accessibility for reproducible research, and local deployment to prevent data leakage when handling sensitive clinical data.
- We designed three prompt variants:
  i. v1: Task description, allowed labels with definitions, classification rules, and output requirements.
  ii. v2: Task description, allowed labels *without* definitions, and output requirements.
  iii. v3: Task description, allowed labels with definitions, and output requirements.
- We tested 0-, 1-, 2-, 4-, 8-, 16-, and 32-shot configurations using the same discharge summary samples across predictions.
- Samples were randomly selected from the training set, with each document containing three to four labels.

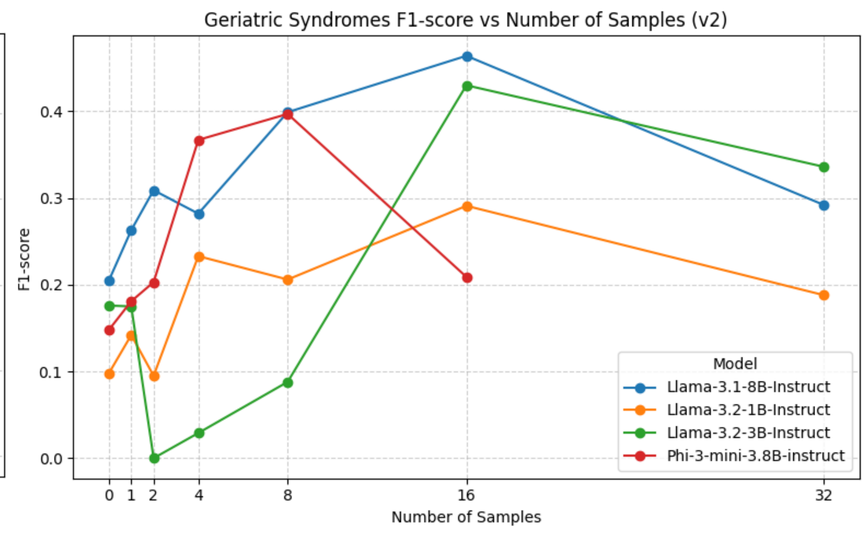
RESULTS

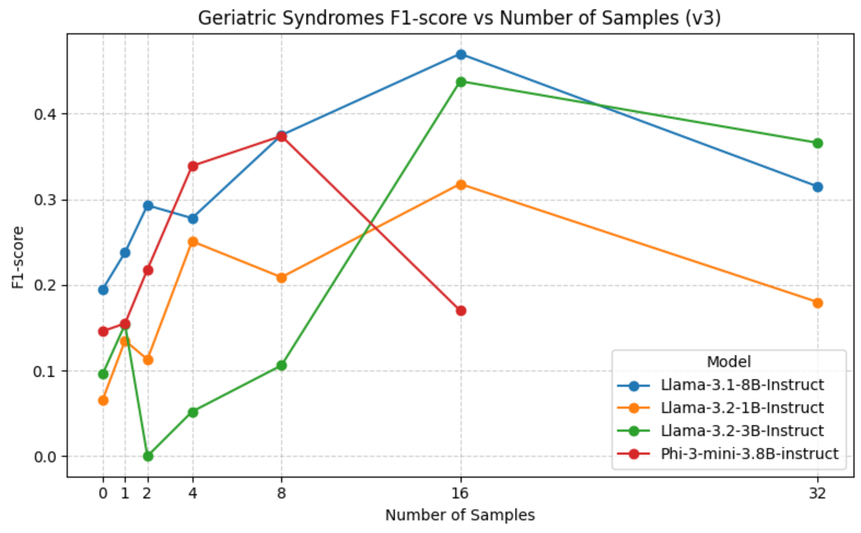












DISCUSSION

- Across the three prompt variants, performance differences were minimal, suggesting that the number of in-context samples contributed more strongly to model performance than prompt design.
- Although all models were reported to support 128k context length, the Phi-3-mini-3.8B-instruct model raised Out-of-Memory errors when processing 32 samples.
- Llama-3.1-8B-Instruct achieved the best overall performance, likely due to its larger parameter size.
- For Adverse Events, the newer and smaller Llama-3.2-3B-Instruct model continued to improve with more samples; however, due to the samples length, 32 samples have nearly approached the 128k context limit.
- Performance tended to increase with the number of in-context samples up to a point, after which further samples led to a decline, suggesting an optimal range for few-shot input size.
- In terms of prediction time, the models ranked from fastest to slowest as follows: Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, and Phi-3-mini-3.8B-instruct.

FUTURE WORK

- Sample selection: We plan to select in-context samples most similar to the target text by computing embedding-based similarity within the training set. This aims to reduce the number of required samples while maintaining or improving performance, potentially lowering prediction time.
- Fine-tuning: Fine-tune the best-performing model from the ICL experiments to improve domain adaptation.
- Model scaling: Evaluate smaller and larger variants of Llama and Phi to assess scalability and efficiency trade-offs in clinical text classification.

[1] Imane Guellil et al. *The First MIMIC Annotated Corpus for the Detection of Geriatric Syndromes and Adverse Events from Discharge Summaries*. In progress.