Automating Infection Indicator Extraction in Home Healthcare through Instruction-Tuned Large Language Models

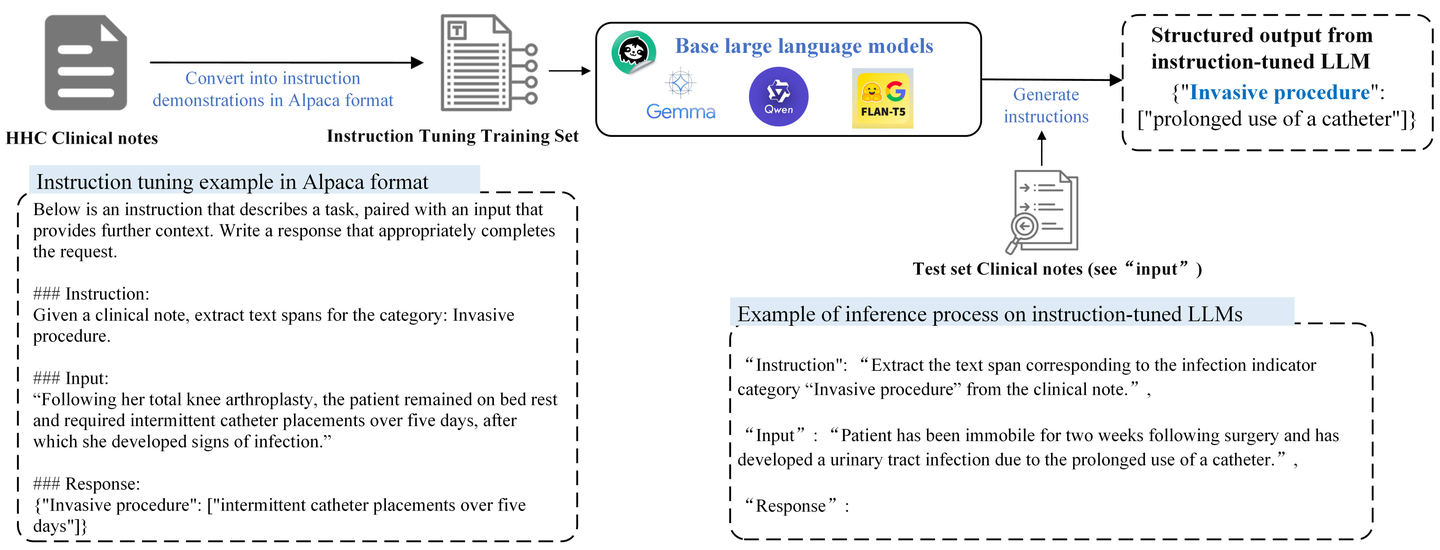
Zidu Xu, PhD, MMed [1], Shuang Zhou, PhD[2], Jiyoun Song, PhD, RN[3], Danielle Scharp, PhD, APRN, FNP-BC[4], Mollie Hobensak, PhD, RN[4], Yan Hu, PhD[5], Jingjing Shang, PhD, RN[1], Maxim Topaz, PhD, RN[1,6,7]

[1] School of Nursing, Columbia University, New York, NY; [2]Department of Surgery, University of Minnesota, Minneapolis, MN; [3]University of Pennsylvania School of Nursing, Philadelphia, PA; [4]Icahn School of Medicine at Mount Sinai, New York, NY; [5]McWilliams School of Biomedical Informatics, University of Texas Health Science at Houston, Houston, TX; [6]VNS Health, New York, NY; [7]Data Science Institute, Columbia University, New York, NY

Contact: zixu2@bwh.harvard.edu

OBJECTIVES: Home healthcare (HHC) clinical notes contain critical infection indicators that clinicians need in structured 'indicator + context' pairs. Data sparsity and limited computing resources hinder automated extraction in decentralized HHC settings. We developed and evaluated a resource-efficient pipeline leveraging instruction-tuned, moderate-sized LLMs and explored a targeted LLM-based data-augmentation strategy to mitigate data scarcity.

Figure 1. Overview of the instruction tuning process



METHODS

Study Data

- 2,145 clinical notes from 717 HHC episodes experienced infection-related hospitalization/ ED visits due to HHC stay (2015-2017, Northeastern US).

Three-Stage Data Augmentation

- ***Manual Annotation***: Targeted enrichment of real notes for rare categories.
- ***Context Mutation:*** Paraphrased surrounding text to enhance data diversity.
- ***Synthetic Generation:*** Created entirely new instances.
- Data quality control through ***automatic filters*** and ***manual validation***.

Instruction Tuning Through QLoRA

- Primary Models: Moderate-sized LLMs (**Gemma-12B, Qwen-14B**).
- Baselines: a prompted Larger-sized LLM (Llama-3.3-70B); a smaller fully-tuned language model (Flan-T5-XL, 3B).
- Finalized outputs: majority vote ensemble.
- Assessed **accuracy** and **reliability**.

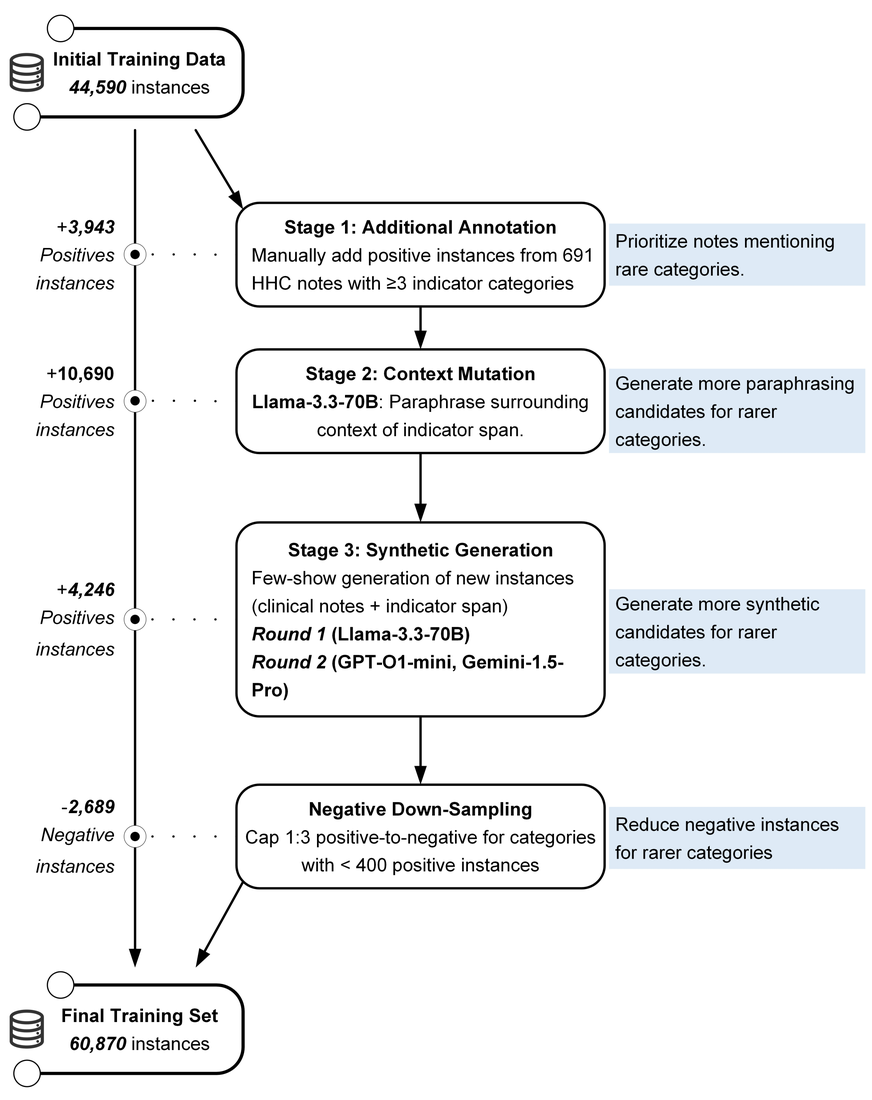


Figure 2. Data augmentation diagram

RESULTS

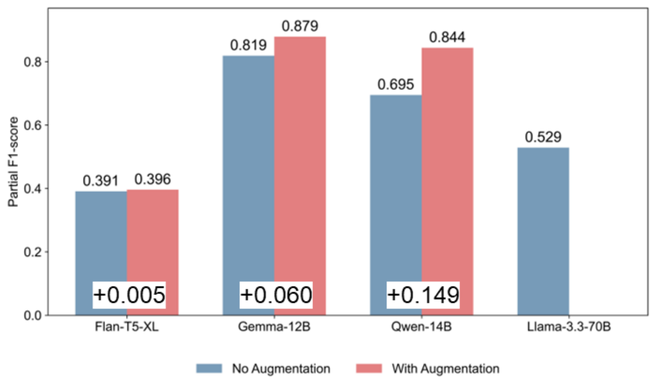


Figure 3. Model performance comparison after majority vote

Best model performances:

- **Augmented Gemma-12B** performed best; partial F1-score: **0.879**.
- Instruction-tuned Gemma-12B & Qwen-14B demonstrated clear **superiority over both baselines**.

Performance varied across infection indicators:

- **Broad, variably-expressed** concepts were more challenging to identify comprehensively (e.g., "Wound Infection Management").
- Categories flagged by **distinct keywords** achieved high accuracy (e.g., "translator" for language barriers).

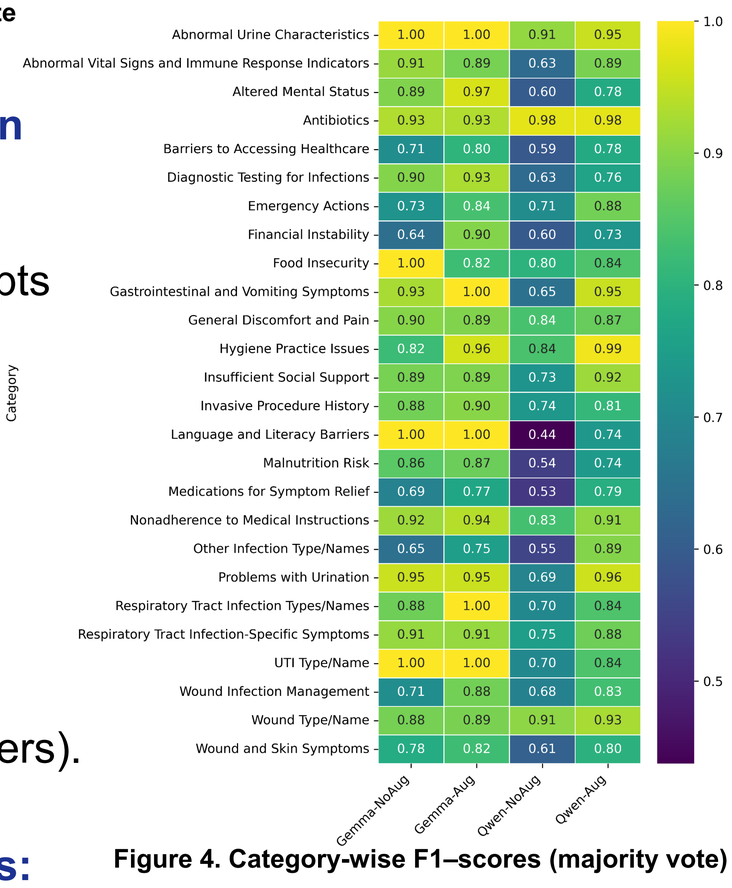


Figure 4. Category-wise F1–scores (majority vote)

Flexible, multi-level clinical insights:

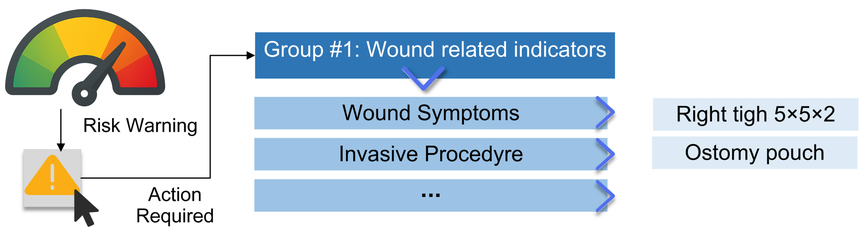


Figure 5. Multi-level support for infection assessment

High-Level: Grouped summaries for quick, at-a-glance assessment.

Granular: Specific indicators with verbatim text for in-depth review and targeted interventions.

Limitations & Future Work

- Refine model to capture full context and complete multi-mention extraction (e.g., identifying all distinct wounds).
- Need ablation studies to optimize augmentation components and improve instruction-tuning dataset quality.