LLMs at Work: Mining NMDOH from Clinical Notes

Reshma Suresh, MS, Jacqueline Naeem, MD, Tuan-Minh Nguyen, PhD, Albert Karam, MS, MBA

Background

- NMDOH = Non-Medical Drivers of Health (e.g. Housing, Food, Transportation, Social Support)
- Rich NMDOH insights lie hidden in free-text clinical notes.
- Structuring this information unlocks population health, risk prediction, and targeted intervention.
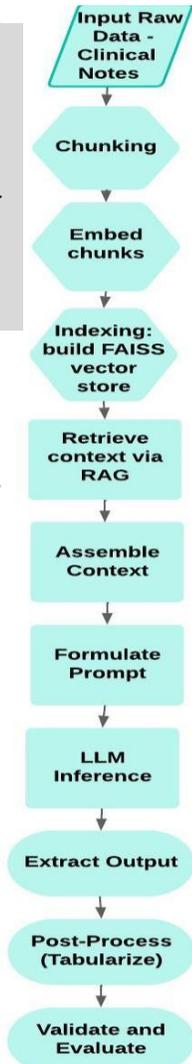

Project Aim

- Develop and validate a Large Language Model (LLM)-based approach to extract NMDOH from clinical notes into structured, tabular data.

Data Setup

Cohort	Granularity	Columns	Response Rate
Adult Patients with Inpatient Admission in 2023-2024	Each row is a unique clinical note (Progress or H&P Note)	NMDOH Categories	Percentage of patients that responded (either Yes or No) to at least one NMDOH category
25544 unique patients	~110K notes	7 categories per note	61%

Approach

```
Input Raw Data - Clinical Notes
        ↓
    Chunking
        ↓
  Embed chunks
        ↓
Indexing: build FAISS vector store
        ↓
Retrieve context via RAG
        ↓
 Assemble Context
        ↓
Formulate Prompt
        ↓
  LLM Inference
        ↓
  Extract Output
        ↓
Post-Process (Tabularize)
        ↓
Validate and Evaluate
```
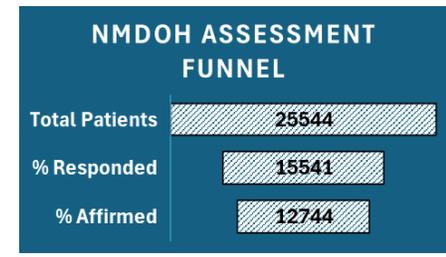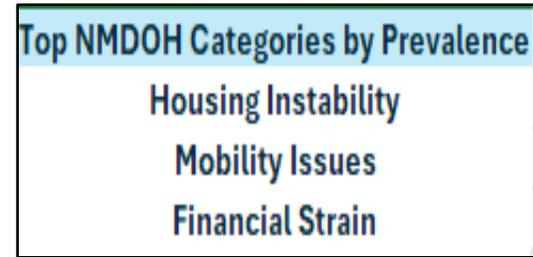
1. Split long notes into manageable chunks so the model/context window isn't exceeded — the code defines a token-aware **chunk_text (tiktoken)** and also uses **RecursiveCharacterTextSplitter(chunk_size=500, overlap=50)** to create document chunks.
2. Encode each text chunk into vector embeddings using **HuggingFaceEmbeddings** (sentence-transformers/all-MiniLM-L6-v2) so semantic similarity search can operate on meaning, not surface text.
3. Load the chunked documents into a FAISS index (**FAISS.from_documents**) to enable fast nearest-neighbor similarity searches over the note content.
4. Perform a similarity search (example query: "What are the patient's social needs?", k=5) to retrieve the most relevant chunks that contain NMDOH evidence.
5. Concatenate the retrieved chunks into a compact context_text that provides the model focused evidence from the note.

Combine the **base_prompt** with the assembled **context_text** (and the patient note when needed) to form **full_prompt** — the single, constrained instruction the LLM will answer.

Call the LLM with the chosen model, temperature, and token limits to produce the extracted NMDOH text.

Results

Top NMDOH Categories by Prevalence
- Housing Instability
- Mobility Issues
- Financial Strain

NMDOH ASSESSMENT FUNNEL	
Total Patients	25544
% Responded	15541
% Affirmed	12744

NMDOH data from Notes vs Z-Code based NMDOH documentation

	Z-Code: Yes	Z-Code: No
Notes: Yes	4118 (TP)	8670 (FN)
Notes: No	616 (FP)	2137 (TN)

Conclusions and Implications

- Sparse documentation via Z-code hides patient needs; NMDOH information from clinical notes can fill the gap.
- Combining NMDOH notes and Z-codes can power predictive models to identify patients at risk for hospital readmission or poor chronic disease outcomes, enabling targeted interventions.
- This integrated data can trigger clinical alerts for unmet social needs, guide care coordination, and inform population health strategies to improve equity.