Applying Generative AI
for
Synthetic Healthcare Data Analysis

Naveen Kilaru

Healthcare issue

Challenge:

Synthesizing patient condition and treatment patterns for personalized risk analysis.

Goal:

Use LLMs to generate condition summaries and predict treatment trends using synthetic EHR data

Key Takeaway:

Early detection support and therapeutic planning guidance

Medical dataset overviews

- **Data Source:** Synthea synthetic health records
- **Used Tables:**
  - conditions.csv: START, STOP, PATIENT, ENCOUNTER, CODE, DESCRIPTION
  - medications.csv: START, STOP, PATIENT, ENCOUNTER, CODE, DESCRIPTION, BASE_COST, TOTALCOST
- **Join Keys:** PATIENT and ENCOUNTER

Sample data

	PATIENT	ENCOUNTER	VISIT_DATE	DIAGNOSIS	MEDICATION	MEDICATION_COST
170	00732e11-5e4d-37b7-01f8-929a25536862	d66f90b7-2020-4639-0cd5-03e9e8cb61a5	2016-06-29	Stress (finding)	insulin isophane human 70 UNT/ML / insulin r...	3835.44
171	00732e11-5e4d-37b7-01f8-929a25536862	d66f90b7-2020-4639-0cd5-03e9e8cb61a5	2016-06-29	Stress (finding)	Hydrochlorothiazide 25 MG Oral Tablet	1.80
172	00732e11-5e4d-37b7-01f8-929a25536862	d66f90b7-2020-4639-0cd5-03e9e8cb61a5	2016-06-29	Stress (finding)	24 HR tacrolimus 1 MG Extended Release Oral Ta...	519.76
173	00732e11-5e4d-37b7-01f8-929a25536862	d66f90b7-2020-4639-0cd5-03e9e8cb61a5	2016-06-29	Stress (finding)	lisinopril 10 MG Oral Tablet	3.64
174	00732e11-5e4d-37b7-01f8-929a25536862	d66f90b7-2020-4639-0cd5-03e9e8cb61a5	2016-06-29	Stress (finding)	amLODIPine 2.5 MG Oral Tablet	3.64

Prompt engineering

- **Prompt type:** In-context learning with few-shot examples
- **Objective:** Generate patient condition summaries and medication explanations
- **Example Prompt:**

Patient ID: 00732e11-5e4d-37b7-01f8-929a25536862
Encounter ID: d66f90b7-2020-4639-0cd5-03e9e8cb61a5
Visit Date: 2016-06-29
Diagnosis: Stress (finding)
Medications Prescribed:
- insulin isophane human 70 UNT/ML ($3835.44)
- Hydrochlorothiazide 25 MG Oral Tablet ($1.80)
- tacrolimus 1 MG Extended Release Oral Tablet ($519.76)
- lisinopril 10 MG Oral Tablet ($3.64)
- amLODIPine 2.5 MG Oral Tablet ($3.64)

Methods employed

- **Model Used:** GPT-4 via OpenAI API
- **Techniques:**
  - Few-shot learning
  - In-context learning
  - Chain-of-thought prompting
- **Data joined and cleaned with Pandas**

Example LLM output from LLM input

Patient **00732e11-5e4d-37b7-01f8-929a25536862** had Stress during encounter **d66f90b7-2020-4639-0cd5-03e9e8cb61a5**.They were prescribed insulin isophane and other medications.

LLM Output: "The patient was diagnosed with Stress (finding) and was prescribed insulin isophane, Hydrochlorothiazide, tacrolimus, lisinopril, and amLODIPine. This regimen suggests a multifactorial management approach for metabolic and cardiovascular concerns."

Evaluation of results

- **Quality Check:**
  - Coherence: ✅
  - Medical terminology use: ✅
  - Summarization relevance: ✅
- **Limitations:**
  - Context sensitivity varies by input structure
  - Limited understanding of progression across encounters

Ideas for improvement

- Incorporate additional tables (e.g., observations, procedures)
- Add reasoning steps through chain-of-thought techniques
- Try disease progression tracking across timelines
- Evaluate multi-modal outputs (e.g., timeline graphs)

Conclusion

- Generative AI offers powerful potential to extract insights from EHR data.
- With structured prompting and integrated datasets, LLMs can aid clinical decision support in a synthetic, privacy-safe way.
- Future work: scale to full patient journeys and add multimodal context.