**🧾 Evaluation Report: Conversational Agent on Health Data**

**📌 Objective**

To build a multi-turn question-answering (QA) agent capable of interacting with structured health datasets and accurately responding to questions involving patient demographics (e.g., race), exposures, and diagnoses. The goal is to emulate an intelligent, memory-aware assistant that can reason over tabular data and maintain context across queries.

**🛠️ Methodology**

**System Overview**

* **LLM**: Azure OpenAI GPT-4 via LangChain
* **Tools**:
  + TableLookup: Filters and counts rows from structured tables (CSV)
  + ConversationMemory: Tracks user queries for context-aware reasoning
* **Datasets**: CSV files including:
  + demographics.csv: contains patient race, age, gender, etc.
  + Additional tables (e.g., exposures.csv, diagnosis.csv) not used in this test
* **Evaluation Framework**:
  + Queries + expected answers (gold labels)
  + Metrics: substring match, BLEU score, semantic similarity

**Evaluation Process**

1. Execute agent on a curated set of natural language queries.
2. Log predictions and compare with gold answers.
3. Compute:
   * ✅ Substring match (binary)
   * 🎯 BLEU Score (n-gram overlap)
   * 🔗 Semantic Similarity (embedding cosine similarity)

**📊 Results**

| **Question** | **Expected Answer** | **Agent Prediction** | **Exact Match** | **BLEU** | **Semantic Similarity** |
| --- | --- | --- | --- | --- | --- |
| How many Asian patients in the dataset? | There are 627 Asian patients | There are 627 Asian patients in the dataset. | ✅ | 1.00 | 1.00 |
| How many White patients in the dataset? | There are 300 White patients | There are 7571 white patients in the dataset. | ❌ | 0.66 | 0.78 |
| How many Hispanic patients in the dataset? | There are 400 Hispanic patients | It appears there are no reported Hispanic ... | ❌ | 0.32 | 0.57 |
| What race was the last one I asked about? | You last asked about Hispanic patients | The last one you asked about was Hispanic. | ❌ | 0.07 | 0.57 |

**✅ Metrics Summary**

| **Metric** | **Value** |
| --- | --- |
| Exact Match Accuracy | 25% |
| Average BLEU Score | 0.51 |
| Avg. Semantic Similarity | 0.73 |

**🧠 Findings**

**✅ Strengths**

* **Correct on first query**: The agent handled structured lookup correctly for the first demographic query.
* **Memory partially functional**: Correctly identified the last race asked as "Hispanic" despite low BLEU (due to phrasing differences).

**❌ Limitations**

* **Lookup errors**: Significant mismatch in counts for White and Hispanic patients suggests the tool may be:
  + Accessing the wrong column
  + Failing to match case or category properly
  + Misreading dataset headers
* **Case sensitivity / fuzzy matching**: “white” vs “White” may affect row filtering.
* **Memory evaluation misleading**: BLEU penalizes phrasing variation even when meaning is intact. Semantic similarity is more reliable for conversational memory tasks.
* **No grounding validation**: Agent doesn’t currently validate output against filtered table contents before responding.

**🔧 Recommendations**

1. **Tool Improvements**
   * Normalize race strings to lowercase before filtering (e.g., row['race'].lower() == 'white')
   * Add fuzzy matching or synonym mapping (e.g., “Caucasian” → “White”)
2. **Evaluation Enhancements**
   * Introduce classification-style metrics: Precision, Recall, F1 for multi-label queries
   * Weight semantic similarity more heavily for conversational/memory tasks
3. **Data Validation**
   * Pre-load and log unique values in each column to catch missing/zero-count races like Hispanic
4. **Memory Trace Debugging**
   * Log conversation memory explicitly to debug failed coreference resolution

**📦 Conclusion**

The health QA agent demonstrates early success in structured data reasoning but requires refinements in its table lookup logic and categorical matching robustness. With improved data handling and post-processing, accuracy and reliability for clinical reasoning tasks can be significantly enhanced.

Would you like this exported as a markdown or PDF file for inclusion in your documentation or reporting?

Top of Form

Bottom of Form