LLM-CGM: A Benchmark for Large Language Model-Enabled Querying of Continuous Glucose Monitoring Data for Conversational Diabetes Management

Elizabeth Healey, Isaac Kohane

Presenter: Shovito Barua Soumma

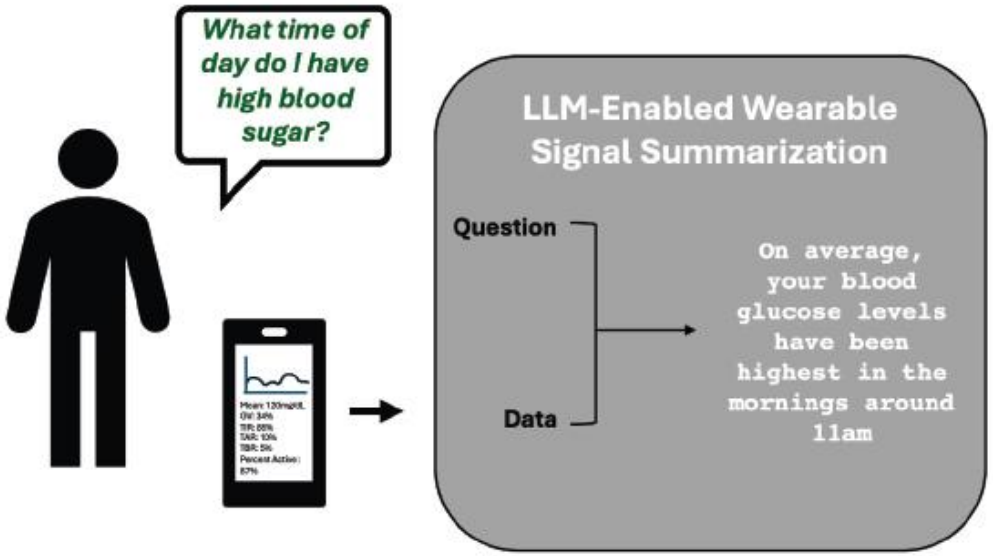
Date: Feb 11, 2025

- Biocomputing 2025: Proceedings of the Pacific Symposium, 2024

- *https://github.com/lizhealey/LLM-CGM*

Summary

- Integration of LLM with CGM data for interpretation
- Bench-mark some q/a tasks (LLM-CGM) for CGM data (4 categories)
  - Performance evaluation of different LLMs using those q/a using synthetic and real cgm data
- Optimizing technique for handling those bench-mark q/a



user could ask a question about their CGM data, and receive a written answer in return, thus transforming the way patients interact with their data

- Q/A tasks can be subject or objective
- Subjective:
  - Is my blood glucose control good?
    - Required patient's medical context (T1 vs T2)
- Focused on objective q/a

Bench Marking Category

- 30 Questions
- inspired by guidelines from the American Diabetes Association (ADA) on glycemic control

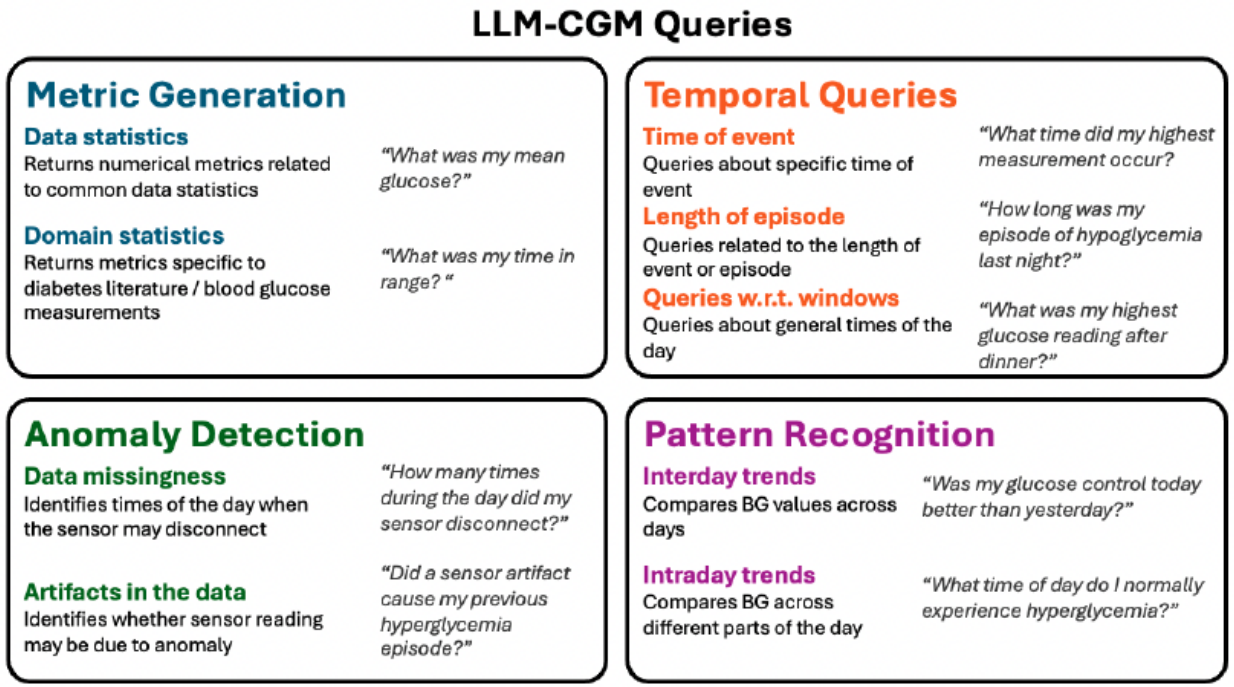


Fig. 2. Benchmarking tasks by category and subcategory

Bench Marking Q/A

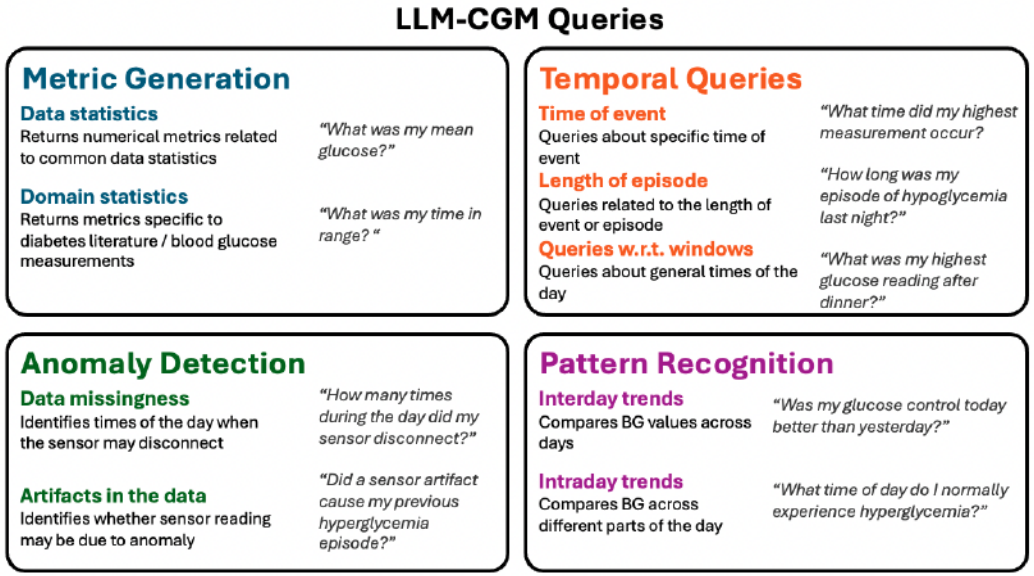


Fig. 2. Benchmarking tasks by category and subcategory

Table 1. LLM-CGM Benchmark Queries and Solutions. The colors correspond to benchmark task categories.

	User Question	Ground Truth Answer
Q1	What was my mean glucose?	Mean of glucose readings
Q2	What was my maximum glucose?	Maximum of glucose readings
Q3	What was the standard deviation of my glucose?	Standard deviation of glucose readings
Q4	What was my minimum glucose?	Minimum of glucose readings
Q5	What was my percent time in range?	Percent time between 70 mg/dL and 180mg/dL
Q6	What was my percent time in hyperglycemia?	Percent time above 180 mg/dL
Q7	What was my percent time in hypoglycemia?	Percent time below 70mg/dL
Q8	What was my glycemic variability?	Standard deviation divided by mean of glucose readings
Q9	What was my percent time in severe hyperglycemia?	Percent of time spent above 250 mg/dL
Q10	What is my estimated A1C?	Using estimated average glucose formula[28]
Q11	What was my percent time in severe hypoglycemia?	Percent time spent below 54 mg/dL
Q12	What time was my blood glucose highest?	Date and time when blood glucose was max
Q13	What day was my glucose control the most out of range?	Day with greatest absolute time outside of range 70-180mg/dL
Q14	What time of the day was my blood glucose lowest?	Date where minimum glucose reached
Q15	When did my most recent episode of hypoglycemia occur?	Time of most recent hypoglycemia episode
Q16	How long was my last episode of hypoglycemia?	Length of most recent period where glucose was consistently below 70mg/dL
Q17	What was my longest time spent in hyperglycemia?	Longest period where glucose was over 180mg/dL
Q18	How many times did I experience hypoglycemia?	Number of episodes where glucose was less than 70mg/dL
Q19	What was my mean overnight blood glucose?	Mean glucose from 12am to 6am**
Q20	What meal of the day did I have the highest blood glucose?	Time window with max glucose where breakfast is 6am-11am, lunch is 11am-4pm, dinner is 5pm-9pm**
Q21	Did I have noctural hypoglycemia?	Yes if blood glucose was less than 70mg/dL between 12am and 6am**
Q22	What was my highest glucose reading during dinner?	Maximum glucose any day between 5pm and 10pm**
Q23	Is there any missingness in the data?	Yes if there are gaps between data longer than 5 minutes
Q24	How many times did my sensor disconnect ?	Number of gaps greater than 5 minutes
Q25	Was my low blood glucose likely due to sensor error?	Yes if reading less than 70 mg/dL due to sensor anomaly*
Q26	Are there any artifacts in the CGM data?	Yes if there was a sensor anomaly in data causing observed glucose reading*
Q27	Was my glucose control today better than yesterday?	Yes if mean glucose on current day was better than previous day**
Q28	Was my time in range improved this week compared to last week?	Yes if time in range for the most recent week was better than the previous week*
Q29	Was my max glucose lower today than yesterday?	Yes if the maximum glucose on most recent day was lower than the previous day
Q30	Did I spend less time in hypoglycemia this week than last week?	Yes if total minutes in hypoglycemia for the most recent week was less than the previous week*
	*Not included in this evaluation	** May be subjective

Framework

- GPT-4
- **LLM-Text** (Naïve approach where raw CGM data is fed as text to an LLM)
- **LLM-Code** (The LLM generates Python code to analyze the CGM data in 3 steps)
- **LLM-CodeChain** (A more advanced approach using LangChain that iterates and refines answers)

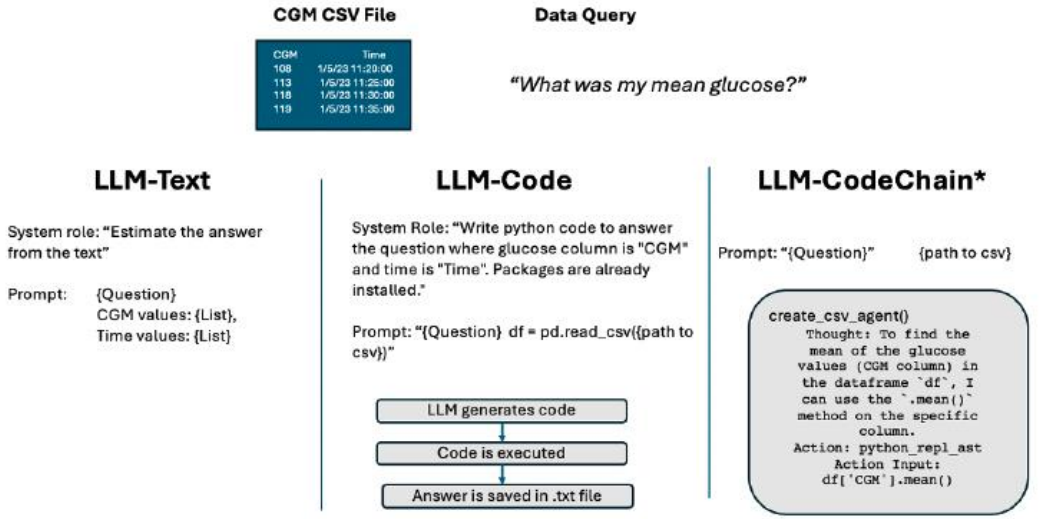


Fig. 3. Model and prompt frameworks included in benchmark for testing and evaluation. LLM-CodeChain leverages builtin functions in Langchain[29]

Retrival augmented generation, where the prompt includes information about diabetes, including definitions of terms and instructions on how to analyze the data

Simulated Data

- *FDA-accepted T1D patient simulator is used*
- N=5
- *Cgm Record/5min*

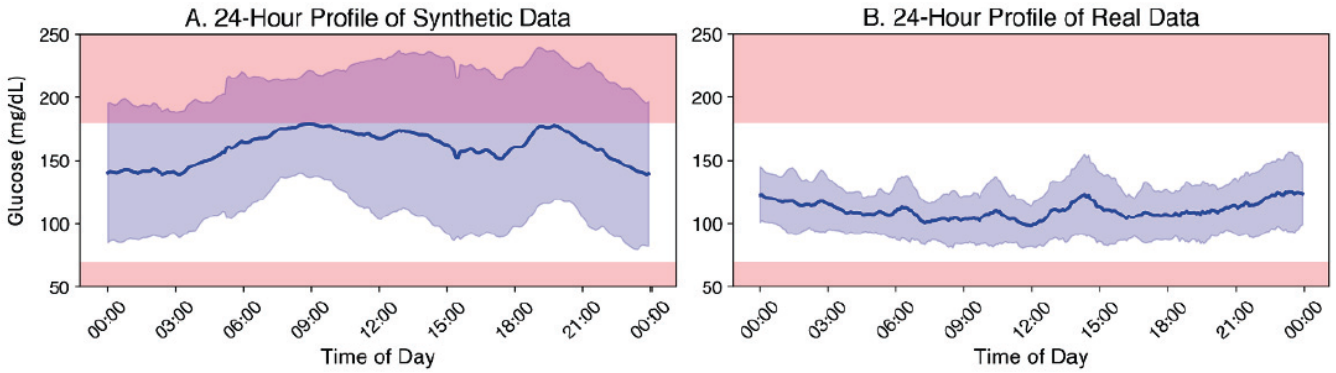


Fig. 4. Data included in benchmark: (A) 24-hour mean and standard deviation of 5 cases from synthetic data simulating patients with T1D. (B) 24-hour mean and standard deviation from 5 cases from the real dataset[33]

- **Mimic real-world glycemic variability**
  - Some simulated patients have **well-controlled glucose levels**.
- Others experience **significant fluctuations** and spend **less than 50% of their time in the healthy glucose range**.

•Real World Data (N=5, 3 Pre Diabetic, 2 Diabetic)

Q8	What was my glycemic variability?	Standard deviation divided by mean of glucose readings

Results

- Categorized by model type and task category
- simpler tasks, such as metric generation, performance was high.
- The more complicated tasks had higher error rates. This was seen through
  - anomaly detection
  - pattern recognition
- LLM-Code > LLM-CodeChain
- What will be the precision of error???

Q17	What was my longest time spent in hyperglycemia?
Q18	How many times did I experience hypoglycemia?
Q19	What was my mean overnight blood glucose?
Q20	What meal of the day did I have the highest blood glucose?
Q21	Did I have noctural hypoglycemia?
Q22	What was my highest glucose reading during dinner?
Q23	Is there any missingness in the data?
Q24	How many times did my sensor disconnect ?
Q25	Was my low blood glucose likely due to sensor error?
Q26	Are there any artifacts in the CGM data?
Q27	Was my glucose control today better than yesterday?
Q28	Was my time in range improved this week compared to last week?
Q29	Was my max glucose lower today than yesterday?

Table 3. Table shows the fraction of CGM cases with correct answer for each question. Results are broken down by the model framework used (LLM-Code vs LLM-CodeChain) and the data type

Metric Generation	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**
LLM-Code Synth (n=5)	1	1	1	1	.8	.8	.8	0	1	1	1
LLM-Code Real (n=5)	1	.8	1	1	1	1	1	0	1	1	1
LLM-Code Total (n=10)	1	.9	1	1	.9	.9	.9	0	1	1	1
LLM-Codechain Synth (n=5)	1	1	1	1	.2	1	1	0	1	1	1
LLM-Codechain Real (n=5)	1	1	1	1	0	1	1	0	1	1	1
LLM-Codechain Total (n=10)	1	1	1	1	.1	1	1	0	1	1	1
Temporal Queries	**Q12**	**Q13**	**Q14**	**Q15**	**Q16**	**Q17**	**Q18**	**Q19**	**Q20**	**Q21**	**Q22**
LLM-Code Synth (n=5)	1	1	.8	.8	.6	0	.2	.8	.4	1	.6
LLM-Code Real (n=5)	1	1	.8	.6	.4	.8	.4	.8	.2	1	.6
LLM-Code Total (n=10)	1	1	.8	.7	.5	.4	.3	.8	.3	1	.6
LLM-Codechain Synth (n=5)	1	0	1	.8	.2	.2	.2	.4	0	.4	0
LLM-Codechain Real (n=5)	1	.2	1	1	.2	.2	.4	.6	.8	1	0
LLM-Codechain Total(n=10)	1	.1	1	.9	.2	.2	.3	.5	.4	.7	0

Anomaly Detection	**Q23**	**Q24**		**Pattern Recognition**	**Q27**	**Q29**
LLM-Code Synth (n=5)	.8	1		LLM-Code Synth (n=5)	0	.4
LLM-Code Real (n=5)	0	0		LLM-Code Real (n=5)	0	.6
LLM-Code Total (n=10)	.4	.5		LLM-Code Total (n=10)	0	.5
LLM-Codechain Synth (n=5)	.8	.2		LLM-Codechain Synth (n=5)	0	0
LLM-Codechain Real (n=5)	0	0		LLM-Codechain Real (n=5)	.4	.4
LLM-Codechain Total(n=10)	.4	.1		LLM-Codechain Total(n=10)	.2	.2