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

Elizabeth Healey[†]
Program in Health, Sciences, and Technology, Massachusetts Institute of Technology, Cambridge, MA 02138, USA
[†]*E-mail: ehealey@mit.edu*

Isaac Kohane
Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

Over the past decade, wearable technology has dramatically changed how patients manage chronic diseases. The widespread availability of on-body sensors, such as heart rate monitors and continuous glucose monitoring (CGM) sensors, has allowed patients to have real-time data about their health. Most of these data are readily available on patients' smartphone applications, where patients can view their current and retrospective data. For patients with diabetes, CGM has transformed how their disease is managed. Many sensor devices interface with smartphones to display charts, metrics, and alerts. However, these metrics and plots may be challenging for some patients to interpret. In this work, we explore how large language models (LLMs) can be used to answer questions about CGM data. We produce an open-source benchmark of time-series question-answering tasks for CGM data in diabetes management. We evaluate different LLM frameworks to provide a performance benchmark. Lastly, we highlight the need for more research on how to optimize LLM frameworks to best handle questions about wearable data. Our benchmark is publicly available for future use and development. While this benchmark is specifically designed for diabetes care, our model implementation and several of the statistical tasks can be extended to other wearable device domains.

Keywords: Large Language Models, Human-AI Interaction, Diabetes, Time Series

1. Introduction

Large language models (LLMs) have demonstrated tremendous promise in transforming how information is automatically distilled and extracted. In clinical medicine, there has been much excitement about how LLMs can transform the way doctors and patients interact with healthcare systems.[1–4] Recent literature has demonstrated the ability of LLMs to extract medical information and provide clinical summaries,[5–8] even using medical images as input.[9,10] These advances have the potential to dramatically change the way that patients and clinicians interact with medical data. Despite these advances, there has been less focus on how LLMs can be used to extract information from time-series data from patient-owned medical devices.

In diabetes management, patient interpretation and understanding of their data is key to making behavioral modifications. In recent years, the use of wearable continuous glucose monitors (CGMs) for diabetes management has increased.[11] These devices are worn on the body and measure interstitial blood glucose approximately every 5 minutes. These devices allow patients to view both their real-time and retrospective data on their smart devices. The insights gained from CGM data are important for helping patients make behavioral and treatment modifications to manage their diabetes.[12] While several applications exist where patients can view their retrospective data, some patients may find the interpretation of CGM data to be challenging.[13]

In this work, we develop a benchmark of CGM question-answering (QA) tasks: LLM-CGM. Figure 1 shows a schematic of the ideal system for LLM-enabled QA for CGM data. In this setup, the 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.

Our contributions can be summarized as follows:

(1) We outline four categories of tasks for CGM QA. We articulate subtasks describing potential natural language queries about the data for each task. For each subtask, we include sample question queries. The final benchmark contains a total of 30 questions.
(2) We provide a module to get the empirical answer questions in the benchmark from any raw CGM data for evaluation.
(3) We implement three distinct baseline approaches to LLM QA of time-series data and show the performance on the benchmark tasks.
(4) We evaluate our benchmark using synthetic and real CGM data of up to 14 days in length.

LLM-CGM can be accessed at https://github.com/lizhealey/LLM-CGM and can be leveraged to evaluate future iterations of LLMs for diabetes.

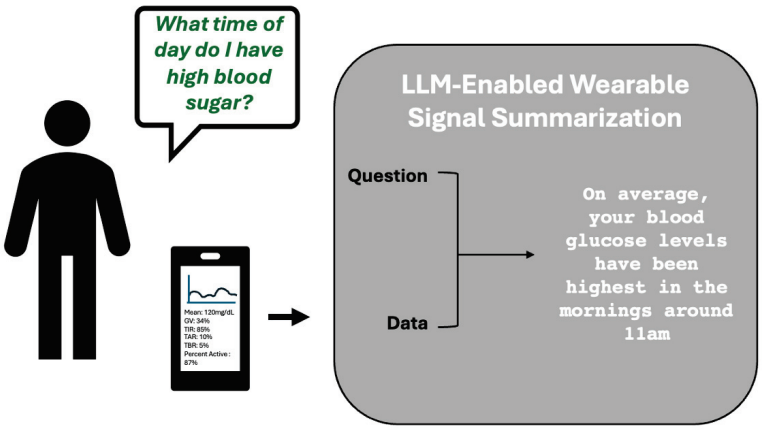

Fig. 1. Illustrative overview.

2. Related Work

2.1. *LLMs of Time Series Interpretation*

Recent work has investigated using LLMs for time-series data analysis,[14] with a subset of this space focusing on how LLMs can be used to interpret and understand time series data.[15] In the medical domain, there has been interest in building benchmarks for question-answering (QA) tasks for wearable data. ECG-QA provides a QA dataset with a benchmark for 70 questions related to electrocardiogram interpretation.[16] The Personal Health Large Language Model (PH-LLM) was developed to provide insights on sleep and fitness goals from wearable data.[17] Similar work was recently published by Merrill et al., where they proposed a Personal Health Insights Agent (PHIA),[18] which leverages code generation and information retrieval to respond to questions about data from wearable devices, such as step count. Our work builds upon this previous work by providing a benchmark for wearable health data interpretation with tasks specific for CGM data.

2.2. *Diabetes Technology*

Interest has also increased in using LLMs to enhance diabetes management through education and personal coaching.[19,20] A previous randomized control trial investigated using voice-based AI to help patients with T2D manage their insulin,[21] and they found that the AI application benefited patients' glycemic control. Other work has investigated a conversational health agent for patients with diabetes, incorporating carbohydrate information and guidelines.[22] Recently, a few works have investigated using LLMs, such as GPT-4,[23] to summarize CGM data.[24,25] These works have explored how LLMs are capable of interpreting CGM data to produce easily understandable summaries. Given the recent interest in the development of diabetes chatbots, there is a need for further investigation of how to optimize LLMs for the analysis of CGM data. Our work fills this gap by presenting a benchmark for conversational queries about CGM data and a preliminary evaluation of different LLM frameworks.

3. Methods

3.1. *Benchmarking Tasks*

Queries of CGM data can have either objective or subjective answers. Many QA tasks for CGM are subjective and depend on the specific patient circumstances. For example, a query of "Is my blood glucose control good?" is subjective and requires consideration of the patient's medical context. In this work, we focus on CGM tasks that can primarily be answered objectively.

Figure 2 gives an overview of the four task categories and subtasks, with example questions. The tasks are broken down into categories that are delineated by both the computational processes required to get an answer and the domain knowledge necessary to understand the task. Many of the questions are inspired by guidelines from the American Diabetes Association (ADA) on glycemic control[26] and current frameworks for analyzing CGM data.[27] Table 1 shows the 30 questions included in this benchmark that are distributed across the task categories. While there are many more types of questions that patients may want to ask about their data, the purpose of these 30 questions was to provide a foundational baseline for a range of query types.

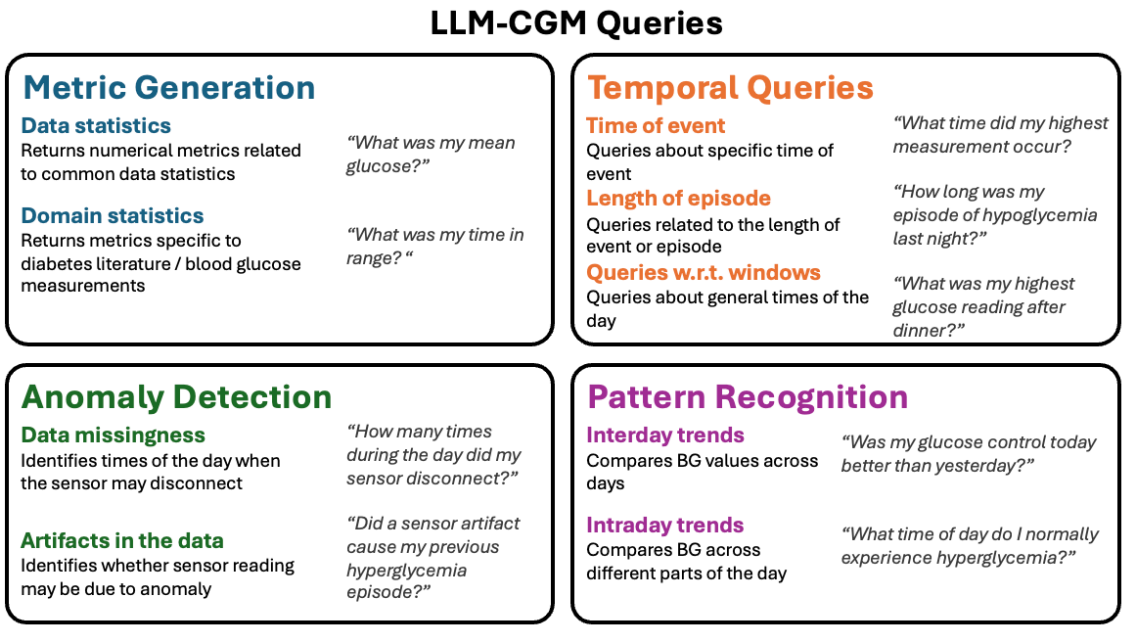

Fig. 2. Benchmarking tasks by category and subcategory

3.2. *Task Evaluation*

Our curated list of 30 tasks has Python-generated solutions. We include the full list of tasks and how they are evaluated in Table 1. Given any comma-separated value (CSV) file as input with a column for the CGM values and timestamp, we automatically compute the answers to the queries using the definitions in the table. For some tasks, the quantitative answer can be subjective. For example, some of the questions depend on the period in which breakfast and dinner are defined. These queries are noted in the table.

3.3. *Model Framework*

In our analysis, we use GPT-4[23] to generate text responses. We test three different frameworks designed to analyze CGM data using GPT-4 that serve as baselines. The details for each model and prompt framework can be found in Figure 3 and we also describe each below.

(1) **LLM-Text:** LLM-Text is a naive implementation where the CGM data and time stamps are inputted to the language model as text as part of the prompt.

(2) **LLM-Code:** LLM-Code is a framework implemented in Python with three main steps. This framework was inspired by recent work examining the ability of GPT-4 to analyze data.[30] In their work, they create a framework where the language model writes code that is automatically executed. We adapt that approach to our setting. In the first step, the

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

LLM writes a Python script that begins by loading a CSV file with the CGM data. We then program LLM-Code to automatically execute the Python script and produce text in a new file. The final answer is obtained from the text file.

(3) **LLM-CodeChain:** Our workflow leverages the *create_csv_agent*() from Langchain[29] that allows the use of a Python tool. This allows the agent to write and run code to analyze the CSV file. We use Langchain to connect to OpenAI's GPT-4 model.[23] The agent takes the preprocessed CSV file as input, along with a prompt. The output is a generated narrative and the log of computations. This framework is most similar to recent work PHIA,[18] where the LLM can iterate through a thought-action chain.

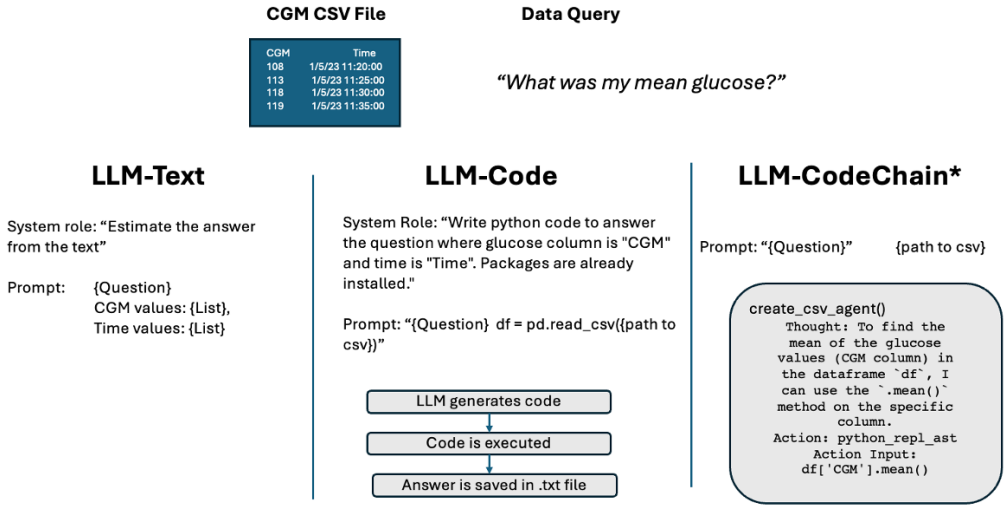

Fig. 3. Model and prompt frameworks included in benchmark for testing and evaluation. LLM-CodeChain leverages builtin functions in Langchain[29]

Prompts: Figure 3 shows the prompts used as input for each of the model frameworks. The prompts always include the query and, depending on the model framework, some information about the context of the data. Future evaluations should include retrieval-augmented generation, where the prompt includes information about diabetes, including definitions of terms and instructions on how to analyze the data.

Technical Specifications: For all model implementations, we generate text using OpenAI's GPT-4.[23] Our repository enables the testing of multiple models; however, for this paper, all experiments were done using the model "gpt-4-0125-preview", with the temperature set to .1.

3.4. *Simulated Data*

While there are many available datasets with CGM data from T1D, many require a data-use agreement to be signed. Since uploading data to open-source LLMs conflicts with the terms of these agreements, we curated our own CGM dataset using an FDA-accepted T1D patient simulator.[31] We generated five different cases of roughly 14 days of CGM data sampled every five minutes. The simulator used was generated from an open-source Python patient simulator.[32] The characteristics of this dataset are visualized in Table 2 and Figure 4. By using the patient simulator, we were able to curate a dataset with variable glycemic control. Simulated cases had significantly varying glycemic signatures and characteristics, with some patients spending a majority of their time in healthy glucose range, and with some individuals spending less than 50% of the time in healthy glucose range.

3.5. *Real Data*

We also used publicly available real CGM data,[33] that was collected from individuals with diabetes, pre-diabetes, and no diabetes. For this work, we only use five individuals in our analysis to demonstrate the performance of LLMs on various CGM QA tasks. This subset included three individuals with pre-diabetes and two without diabetes. The characteristics of this dataset can be visualized in Table 2 and Figure 4.

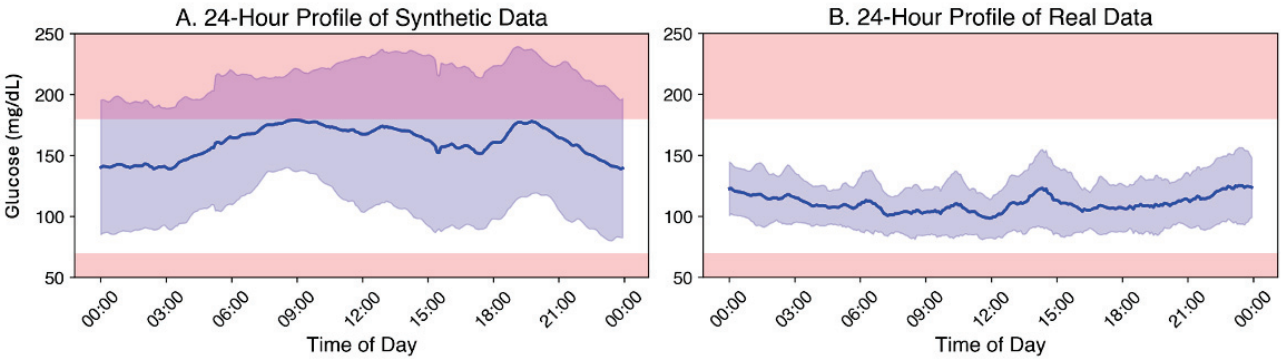

Fig. 4. Data included in benchmark: (A) 24-hour mean and standard deviation of 5 cases from synthetic data simulating patients with T1D. (B) 24-hour mean and standard deviation from 5 cases from the real dataset[33]

Table 2. Characteristics of data. We show the mean value for each of the statistics, as well as the minimum value in the dataset and maximum value in the dataset.

	Synthetic T1D Data (n=5)			Real Data (n=5)		
	Mean	Min	Max	Mean	Min	Max
Number of data points	4033.0 (0.0)	4033	4033	1875.2 (171.814)	1779	2180
Average glucose (mg/dL)	168.085 (25.887)	130.298	196.627	108.052 (7.021)	97.013	116.556
Glucose management indicator	7.331 (0.619)	6.427	8.013	5.895 (0.168)	5.631	6.098
Coefficient of variation	0.3 (0.04)	0.242	0.354	0.172 (0.037)	0.135	0.225
Minimum glucose (mg/dL)	53.15 (10.846)	43.888	71.121	65.0 (4.528)	58	69
Maximum glucose (mg/dL)	352.354 (63.41)	267.212	400	192.8 (33.937)	144	234
Percent time sensor active	1.0 (0.0)	1	1	0.465 (0.043)	0.441	0.541
Percent time in range (70mg/dL-180mg/dL)	0.644 (0.175)	0.472	0.901	0.987 (0.011)	0.975	0.997
Percent time above range 1 (>180mg/dL)	0.349 (0.177)	0.099	0.526	0.006 (0.008)	0	0.02
Percent time above range 2 (>250mg/dL)	0.094 (0.087)	0.001	0.224	0.0 (0.0)	0	0
Percent time below range 1 (<70mg/dL)	0.007 (0.011)	0	0.026	0.004 (0.005)	0.001	0.012
Percent time below range 2 (<54mg/dL)	0.001 (0.002)	0	0.005	0.0 (0.0)	0	0

4. Results

In Table 3, we show the results categorized by the model type and the task categories. The questions are shown individually across all cases, with total scores also listed for each task category. We found that for simpler tasks, such as metric generation, performance was high. Errors were often caused by a misinterpretation of the task. For example, when computing

glycemic variability, the LLM would return the standard deviation, not the coefficient of variation (Q8). The more complicated tasks had higher error rates. This was seen through anomaly detection tasks and pattern recognition tasks. We also note that the performance of LLM-Code compared to LLM-CodeChain varied depending on the tasks.

Table 4 gives examples of incorrect answers by framework. During our evaluation, there were many times when the model did not produce an answer. This was often due to an error in the original code. For these instances, instead of rerunning the example, we counted the instances as inaccurate. These instances often occurred for tasks that were complicated, and the model output suggested the limitation was due to inadequate information. For most tasks, LLM-Code outperformed LLM-CodeChain. A notable limitation with LLM-Code is that code is only written once, so the agent has no ability to rewrite code based on the output. This is seen as an example in Table 4 where the length of the most recent episode of hypoglycemia was not able to be computed. However, for some of the more complicated temporal queries, LLM-CodeChain outperforms LLM-Code for the real cases.

Performance for the anomaly detection tasks and pattern recognition tasks were particularly low. This was due to the fact that the computations necessary to answer these was more complicated than to those of the other tasks. Without any information in the prompt about what to execute, the LLM fails to answer correctly most of the time. Additionally, the prompts did not include any information on what day "today" was, impairing the performance.

We do not show the results for the LLM-Text framework due to the fact that there was very poor performance for most of the tasks. The data used in our evaluation had CGM traces of up to 14 days in length. This caused the token size of the model input to be extremely large and the LLM struggled to return even basic estimates. We expect that the performance could likely increase with smaller amounts of CGM data. An example output of LLM-Text to Q1 is seen in Table 4.

There was some subjectivity when grading whether or not the LLM outcome was accurate. For example, some numerical results were rounded, or within a very small margin of error. For questions that returned percentages and values, answers were marked as correct if they were equivalent when rounded to the nearest whole number. For questions related to meal times, such as Q19 and Q22, answers were marked correct if they were within 10mg/dL of the solution. We omitted four questions in the analysis presented in the paper. We omitted Q28 and Q30 since they are dependent on how a week is defined. We also omitted Q25 and Q26 because the data we used had no documented artifacts.

5. Conclusions

In this work, we developed a benchmark for LLM-enabled CGM QA tasks. We hope that this work promotes further investigation of conversational agents for diabetes management. Our work highlighted the potential for innovation of LLM frameworks for wearable data analysis. LLM-Code and LLM-CodeChain both involved leveraging Python to analyze the data based on the LLM output. LLM-Code was limited by the fact that it was designed only to be able to write one Python script. We suspect for more complex tasks, LLM-CodeChain has benefits that should be further investigated.

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	

Table 4. Error Examples

Framework	Error Examples
LLM-Text	Q1: "Given the extensive list of values, a precise calculation here isn't feasible without computational tools. However, we can provide a rough estimate by observing the range and distribution of values. The glucose values range from lows around 137 to highs of 322. Most values seem to cluster in the mid-100s to mid-200s, with occasional peaks above 300. Given this distribution, it's reasonable to estimate that the mean glucose level might fall somewhere in the mid to upper 100s, possibly around the 180 to 220 range."
LLM- Code	Q17: "Based on the data provided, your longest time spent in hyperglycemia was 23 hours and 35 minutes." Q16: "If your episode of hypoglycemia is ongoing, it's not possible to determine its duration until it has concluded."
LLM- CodeChain	Q20: "The period of the day when you had the highest blood glucose was at 07:45:00 on January 9, 2024" Q24: "Without further information on how sensor disconnections are indicated in the data, it's not possible to determine the number of times the sensor disconnected based on the provided information."

This study had several limitations. We used a general purpose model that had not been fine-tuned on any diabetes guidelines as our baseline. In our work, we showed baselines for the performance of GPT-4 in answering these questions. Future work should investigate different models, as well as different prompting techniques. Future work should also investigate performance on different data. In this work, we used a mix of synthetic data and real data. The performance of these frameworks may vary with real CGM data that is different than what was tested. In particular, the performance may vary based on the length of the data being analyzed.

There are particular safety concerns when developing and implementing LLMs for diabetes management. Even in the absence of LLM-generated medical advice, incorrect assessment of glucose data could cause patients to incorrectly dose insulin and put them at risk for life-threatening hypoglycemia. While the framework we proposed in this work is a promising research direction, incorrect answers pose a safety risk. These safety risks should inform future model development and evaluation. Lastly, future work should explore how clinicians and patients evaluate the output of these LLMs. While this work focused on benchmarking the accuracy of QA tasks, there is much to be investigated to determine the clinical utility of LLM-

enabled CGM analysis. The 30 questions in this benchmark were included to demonstrate the breadth of questions that could be asked about CGM data. In the future, the benchmark will expand with questions derived from patients themselves.

6. Acknowledgments

This work was supported by the National Science Foundation Graduate Research Fellowship Program under grant number 2141064.

References

1. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan and D. S. W. Ting, Large language models in medicine, *Nat. Med.* **29**, 1930 (August 2023).
2. N. H. Shah, D. Entwistle and M. A. Pfeffer, Creation and adoption of large language models in medicine, *JAMA* **330**, 866 (September 2023).
3. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera Y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam and V. Natarajan, Large language models encode clinical knowledge, *Nature* **620**, 172 (August 2023).
4. T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, N. Tomasev, S. Azizi, K. Singhal, Y. Cheng, L. Hou, A. Webson, K. Kulkarni, S. Sara Mahdavi, C. Semturs, J. Gottweis, J. Barral, K. Chou, G. S. Corrado, Y. Matias, A. Karthikesalingam and V. Natarajan, Towards conversational diagnostic AI (January 2024).
5. D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly and A. S. Chaudhari, Adapted large language models can outperform medical experts in clinical text summarization, *Nat. Med.* **30**, 1134 (April 2024).
6. M. Agrawal, S. Hegselmann, H. Lang, Y. Kim and D. Sontag, Large language models are Few-Shot clinical information extractors (May 2022).
7. L. Tang, Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding, G. Durrett, J. F. Rousseau, C. Weng and Y. Peng, Evaluating large language models on medical evidence summarization, *NPJ Digit Med* **6**, p. 158 (August 2023).
8. K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. O. Sabel, J. Ricke and M. Ingrisch, ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports, *Eur. Radiol.* **34**, 2817 (May 2024).
9. S. Lee, W. J. Kim, J. Chang and J. C. Ye, LLM-CXR: Instruction-Finetuned LLM for CXR image understanding and generation (May 2023).
10. M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, A. V. Parwani, A. Zhang and F. Mahmood, A visual-language foundation model for computational pathology, *Nat. Med.* **30**, 863 (March 2024).
11. G. Cappon, M. Vettoretti, G. Sparacino and A. Facchinetti, Continuous glucose monitoring sensors for diabetes management: A review of technologies and applications, *Diabetes Metab. J.* **43**, 383 (August 2019).
12. N. Ehrhardt and E. Al Zaghal, Continuous glucose monitoring as a behavior modification tool, *Clin. Diabetes* **38**, 126 (April 2020).
13. K. Mackett, H. Gerstein and N. Santesso, Patient perspectives on the ambulatory glucose profile

report for type 1 diabetes management in adults: A national online survey, *Can J Diabetes* **47**, 243 (April 2023).
14. X. Zhang, R. R. Chowdhury, R. K. Gupta and J. Shang, Large language models for time series: A survey, *arXiv [cs.LG]* (February 2024).
15. E. Fons, R. Kaur, S. Palande, Z. Zeng, S. Vyetrenko and T. Balch, Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark, *arXiv [cs.CL]* (April 2024).
16. J. Oh, G. Lee, S. Bae, J.-M. Kwon and E. Choi, ECG-QA: A comprehensive question answering dataset combined with electrocardiogram, *arXiv [q-bio.QM]* (June 2023).
17. J. Cosentino, A. Belyaeva, X. Liu, N. A. Furlotte, Z. Yang, C. Lee, E. Schenck, Y. Patel, J. Cui, L. D. Schneider, R. Bryant, R. G. Gomes, A. Jiang, R. Lee, Y. Liu, J. Perez, J. K. Rogers, C. Speed, S. Tailor, M. Walker, J. Yu, T. Althoff, C. Heneghan, J. Hernandez, M. Malhotra, L. Stern, Y. Matias, G. S. Corrado, S. Patel, S. Shetty, J. Zhan, S. Prabhakara, D. McDuff and C. Y. McLean, Towards a personal health large language model (June 2024).
18. M. A. Merrill, A. Paruchuri, N. Rezaei, G. Kovacs, J. Perez, Y. Liu, E. Schenck, N. Hammerquist, J. Sunshine, S. Tailor, K. Ayush, H.-W. Su, Q. He, C. Y. McLean, M. Malhotra, S. Patel, J. Zhan, T. Althoff, D. McDuff and X. Liu, Transforming wearable data into health insights using large language model agents (June 2024).
19. G. G. R. Sng, J. Y. M. Tung, D. Y. Z. Lim and Y. M. Bee, Potential and pitfalls of ChatGPT and Natural-Language artificial intelligence models for diabetes education, *Diabetes Care* **46**, e103 (March 2023).
20. B. Sheng, Z. Guan, L.-L. Lim, Z. Jiang, N. Mathioudakis, J. Li, R. Liu, Y. Bao, Y. M. Bee, Y.-X. Wang, Y. Zheng, G. S. W. Tan, H. Ji, J. Car, H. Wang, D. C. Klonoff, H. Li, Y.-C. Tham, T. Y. Wong and W. Jia, Large language models for diabetes care: Potentials and prospects, *Sci Bull (Beijing)* **69**, 583 (March 2024).
21. A. Nayak, S. Vakili, K. Nayak, M. Nikolov, M. Chiu, P. Sosseinheimer, S. Talamantes, S. Testa, S. Palanisamy, V. Giri and Others, Use of Voice-Based conversational artificial intelligence for basal insulin prescription management among patients with type 2 diabetes: A randomized clinical trial, *JAMA Network Open* **6**, e2340232 (2023).
22. M. Abbasian, Z. Yang, E. Khatibi, P. Zhang, N. Nagesh, I. Azimi, R. Jain and A. M. Rahmani, Knowledge-Infused LLM-Powered conversational health agent: A case study for diabetes patients (February 2024).
23. OpenAI, GPT-4 technical report (March 2023).
24. C. Martinez-Cruz, J. F. G. Guerrero, J. L. L. Ruiz, A. J. Rueda and M. Espinilla, A first approach to the generation of linguistic summaries from glucose sensors using GPT-4, in *Proceedings of the 15th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2023)*, (Springer Nature Switzerland, 2023).
25. E. Healey, A. Tan, K. Flint, J. Ruiz and I. Kohane, Leveraging large language models to analyze continuous glucose monitoring data: A case study, *medRxiv* (April 2024).
26. G. Assessment, 6. glycemic targets: standards of medical care in diabetes—2022, *Diabetes Care* **45**, p. S83 (2022).
27. L. Czupryniak, G. Dzida, P. Fichna, P. Jarosz-Chobot, J. Gumprecht, T. Klupa, M. Mysliwiec, A. Szadkowska, D. Bomba-Opon, K. Czajkowski, M. T. Malecki and D. A. Zozulinska-Ziolkiewicz, Ambulatory glucose profile (AGP) report in daily care of patients with diabetes: Practical tips and recommendations, *Diabetes Ther.* **13**, 811 (April 2022).
28. D. M. Nathan, J. Kuenen, R. Borg, H. Zheng, D. Schoenfeld, R. J. Heine and A1c-Derived Average Glucose Study Group, Translating the A1C assay into estimated average glucose values, *Diabetes Care* **31**, 1473 (August 2008).
29. H. Chase, "langchain-experimental 0.0.40" (2023), Version 0.0.40, Software available from

https://pypi.org/project/langchain-experimental/0.0.40/.
30. L. Cheng, X. Li and L. Bing, Is GPT-4 a good data analyst?, *arXiv [cs.CL]* (May 2023).
31. C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev and C. Cobelli, The UVA/PADOVA type 1 diabetes simulator: New features, *J. Diabetes Sci. Technol.* **8**, 26 (January 2014).
32. J. Xie, Simglucose v0.2.1 . (2018) [Online]. Available: https://github.com/jxx123/simglucose.
33. H. Hall, D. Perelman, A. Breschi, P. Limcaoco, R. Kellogg, T. McLaughlin and M. Snyder, Glucotypes reveal new patterns of glucose dysregulation, *PLoS Biol.* **16**, p. e2005143 (July 2018).