Evaluating the Zero-Shot Predictive Ability of Large Language Models for Continuous Glucose Monitoring Data

by

Junyan Tan (Samuel)

A thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science
(Honors in Data Science, Department of Statistics)
at the University of Michigan

Thesis Advisor:

Irina Gaynanova, Associate Professor of Biostatistics

Mentor:

Renat Sergazinov, PhD

April 22, 2025

DEDICATION

To my parents, who have never doubted me—even in my darkest moments, when everything felt hopeless and I stopped believing in myself. They have supported me unconditionally to this very day. I haven't always been the easiest child, but my mother has always welcomed me with open arms. She walked beside me through my hardest times, and I truly couldn't have made it here without her. My father taught me how to be a good human being—how to think critically and independently. He is a remarkable businessman, and an even greater father.

To Charlotte Xu—thank you for being a wonderful friend and companion throughout the final year of my undergraduate life, both in and out of school. We've shared memories I will never forget, and writing this thesis would have been far more difficult and lonely without you.

To Neo Kok—thank you for being my research buddy. I truly cherish the time we spent with Charlotte working through CGM data and iglu. Your presence made the process much more enjoyable.

To Soren Shao—thank you for being my guitarist this past year. Playing in the band with you was exactly the change of pace and sense of calm I needed while writing this thesis.

To Siyuan Zhang (Balm)—you've always struck me as someone who sees things deeply and clearly. Thank you for being there to listen when I felt lost, sad, or unsure. You stood by me in some of my darkest moments, and I'm grateful beyond words.

To Walter Gui, Cyril Liu, and Amber Zeng—hey, I made it! I know times can be tough, and things don't always go as planned, but I'm rooting for each of you. I hope everything works out in the end.

To Gary Li, Leo Guo, and Bill Zhang—thank you for sharing the early years of my college life. The time we spent together in that small town in Southern California may have been some of the best years of my life.

I am, truly, standing on the shoulders of giants.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest appreciation to my thesis advisor, Professor Irina Gaynanova. She introduced me to the world of scientific research and taught me everything she knew. This paper would not have been possible without her unwavering support and guidance throughout the process. I am especially grateful for the invaluable opportunities she has provided me over the past year. It has been a challenging time, and I was often confused and uncertain about my path—thank you for your patience, understanding, and belief in me.

I am also deeply thankful to Renat Sergazinov, who, at the time of writing, is a research scientist at Meta. Renat introduced me to the world of LLM and contributed many brilliant ideas to this project. He generously shared his knowledge, and every conversation with him reflected the depth and brilliance of a truly exceptional mind.

In addition, I would like to thank Professor Ji Zhu of the University of Michigan. He wrote a recommendation letter for my Master's application and has been a source of inspiration in my pursuit of research in statistics. He is a remarkable statistician and someone I have always aspired to become.

I am grateful to Professor Sebastian Zollner, also of the University of Michigan. He taught the most impactful class I have ever taken and offered thoughtful guidance for my research career. His mentorship helped me discover my research interests, especially in the field of statistical genetics.

I would also like to thank the advisors and faculty members in both the Department of Statistics and the Department of Biostatistics—especially Ms. Gina Cornacchia—for their support in course selection and for helping me find research opportunities.

This thesis presents a selection of my research on Large Language Models and their applications in forecasting CGM data. A poster based on the preliminary results of this thesis received the Best Poster Award at the 2025 Michigan Student Symposium for Interdisciplinary Statistical Sciences.

I wrote this thesis as an undergraduate honors student majoring in Data Science and Statistics. Although the thesis adopts "we" as the first-person pronoun, I am the sole author and take full responsibility for any errors that may remain.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

ABSTRACT vi

CHAPTER

1 Introduction **1**

1.1 Background and Motivation 1
1.2 Large Language Model (LLM) 2

2 Literature Review **3**

2.1 Benchmarking and Evaluation of CGM Data Forecasting Models 3
2.2 LLM-based Approaches for Time Series Forecasting 4
2.3 Retrieval-Augmented Generation (RAG): Theory and Applications 4

3 Data **6**

3.1 Data Processing and Characteristics 6
3.2 Data Preprocessing 7
3.2.1 Interpolation & Segmentation 7
3.2.2 Data Splitting 9

4 Method **10**

4.1 Problem Definition 10
4.2 Baseline Models 10
4.3 Direct-Prompt Forecasting Framework for Large Language Models 11
4.3.1 Contextual Prompt Engineering for Forecasting 13
4.4 Retrieval-Augmented Generation Forecasting 14
4.4.1 Selection of Hyperparameters 17
4.5 Evaluation of Model Performance 17

5 Results **19**

5.1 Individual Forecasting Results 19
5.1.1 Context-Based Forecasting Outcomes 19
5.1.2 Scenario-Based Forecasting Outcomes 21
5.2 Aggregate Dataset Results 21

5.2.1 Direct-Prompt Forecasting 21
5.2.2 Retrieval-Augmented Generation 22
5.2.3 Summary of Best-Performing Models 24

6 Discussion 25

6.1 Strengths of the Direct-Prompt Framework 25
6.1.1 Dynamic Adaptation to Contextual Scenarios 25
6.1.2 Accuracy of Direct-Prompt Forecasting 26
6.2 Evaluating Retrieval-Augmented Generation 27
6.3 Model Recommendations and Practical Guidance 28
6.4 Cost-Efficient and Clinically Scalable Forecasting with LLMs 28
6.5 Limitations of the Current Study 29
6.6 Future Research Directions 29

7 Conclusion 31

8 Ethical Considerations and Data Privacy 33

APPENDICES 34

BIBLIOGRAPHY 37

ABSTRACT

With the growing adoption of Continuous Glucose Monitors (CGMs) in clinical settings, accurate blood glucose forecasting has become pivotal for optimal diabetes management. Although traditional statistical models and time series models have been used to predict glucose levels from CGM data, most require extensive training on large CGM datasets and do not account for important patient demographics and prediction context—such as insulin therapy type, diabetes type, and other lifestyle factors. The lack of an easy-to-use, out-of-the-box model that seamlessly integrates demographic information has hindered broader clinical implementation and the timely prediction of glycemic risks. In this work, we address these challenges by repurposing a prompt-based, zero-shot Large Language Model (LLM) framework for CGM data forecasting. Rather than training a specialized regression model, we convert CGM data (with blood glucose level readings at regular intervals) into text-based prompts and augment them with relevant patient demographics. We then query out-of-the-box LLMs, without additional fine-tuning, to predict future blood glucose levels based solely on these text-formatted prompts. We evaluate model performance against conventional approaches (e.g., linear regression) and time series models (e.g. ARIMA), using median absolute error (MAE) and root mean squared error (RMSE) as our primary metrics. We anticipated that by reducing the need for extensive training and simplifying the inclusion of patient characteristics, this LLM-based approach could significantly lower barriers to clinical implementation and advance the personalization of diabetes management.

Our code is available online at https://github.com/Toshihiko-tan/Gluco-LLM.

Keywords: Continuous Glucose Monitoring (CGM), Large Language Model (LLM), Zero-Shot Learning, Time Series Forecasting, Retrieval-Augmented Generation (RAG), Diabetes Technology

CHAPTER 1

Introduction

1.1 Background and Motivation

Continuous Glucose Monitors (CGMs) are medical devices that can provide real-time glucose readings via a small sensor inserted under the skin, a transmitter that attaches to the sensor, and a receiver that display the transmitted readings. Since its commercial debut in 2000, CGMs have gained increasing prominence, particularly among individuals with type 1 diabetes, demonstrating significant benefits in terms of glycemic control (Carlson et al., 2017). Along with its addition in Medicare since the year 2023, the American Diabetes Association started to recommend consistent use of CGMs for people with type 1 diabetes and type 2 diabetes who are on insulin therapy (Oser and Oser, 2024). Additionally, with the advent of increasingly accurate, user-friendly, and cost-effective models on the market, CGMs have also garnered increasing interest in clinical applications (Rodbard, 2016). As the popularity of CGMs continues to rise, there is an urgent need for improved forecasting methods to effectively manage glucose levels for both diabetic individuals and physicians alike. Efforts have been made to establish a benchmark for forecasting Continuous Glucose Monitoring (CGM) data utilizing conventional machine learning algorithms and models specifically designed for time series analysis (Sergazinov et al., 2024). However, none of these methods have been able to incorporate prediction scenarios and context into their predictions. In this thesis, we propose a novel Large Language Model (LLM)-based approach that can account for both demographic information and the daily actions taken by individuals into the forecasting of CGM data. We establish a baseline using a straightforward direct-prompting approach and subsequently enhance the model's accuracy by applying Retrieval-Augmented Generation (RAG).

1.2 Large Language Model (LLM)

Large Language Model (LLM) is a type of machine learning model that is specifically trained and designed for natural language processing (NLP) tasks, especially for language generation. Generative Pre-Trained Transformer (GPT) models (Radford and Narasimhan, 2018), such as GPT-4o and GPT-4o mini, developed by OpenAI, have been extensively studied and utilized in both academic and industrial settings (Zhao et al., 2025). Although initially designed for NLP tasks, there has been a growing interest in employing LLMs for Time-Series Forecasting tasks. In this thesis, we adopt a zero-shot forecasting paradigm, in which an LLM generates predictions for a new time series without any additional fine-tuning or task-specific examples. LLMs have demonstrated surprisingly high accuracy, and their user-friendly interface enables accurate predictions across a diverse range of datasets (Williams et al., 2025). Given that CGM data, in its format of active glycemic readings, constitutes a time series, we have been motivated to implement LLM for the forecasting of CGM data. In this thesis, we will use a pre-trained LLM to forecast CGM data in a zero-shot setting. This means that the pre-trained LLM will generate predictions for a new time series without any additional fine-tuning or training.

What we mean by "zero-shot"

In zero-shot forecasting, the model is presented with a new time series and its context prompt only—no parameter updates on the LLM are performed. Rather than adapting the model to our CGM dataset, we leverage the LLM's pre-trained knowledge and properly formatted prompt instructions to produce predictions. This differs from fine-tuning approaches, where the model is re-trained on domain-specific data. Zero-shot thus tests the LLM's out-of-the-box generalization to a forecasting task.

CHAPTER 2

Literature Review

2.1 Benchmarking and Evaluation of CGM Data Forecasting Models

In response to the growing clinical application of Continuous Glucose Monitoring (CGM) data, there has been a substantial development of CGM analysis software. Prominent examples include iglu (Broll et al., 2021), GlyCulator (Chrzanowski et al., 2022), and cgmanalysis (Vigers et al., 2019). These software programs provided physicians with a comprehensive overview of the glycemic control status of diabetic individuals, including their associated risks. However, research focusing on forecasting CGM data has been limited. Notably, Gecili et al. presented a predictive tool for CGM data using a longitudinal analysis approach on glucose curves (Gecili et al., 2020). Furthermore, Prendin et al. conducted a comparative analysis of linear and non-linear methods for the prediction of CGM data (Prendin et al., 2021).

However, the most recent and comprehensive work on this subject is GlucoBench (Sergazinov et al., 2024), which provides a comprehensive benchmark of CGM data prediction models. One of the primary findings presented in GlucoBench is the absence of a universally optimal model, as different datasets may yield a different most optimal model. GlucoBench investigated the underlying causes of the performance disparity between datasets and attributed it to the dataset's size, population demographics, and diabetic status. Consequently, for our LLM-based approach, the elements we seek to identify and the theories we intend to validate are as follows: to ascertain whether LLMs can effectively address the disparities identified in GlucoBench and whether they can achieve a higher forecasting accuracy compared to the traditional forecasting models implemented in GlucoBench's list. The data formatting and processing code of GlucoBench will serve as the code foundation for this project.

2.2 LLM-based Approaches for Time Series Forecasting

There have been attempts to apply Large Language Models (LLMs) to time series forecasting, with early research primarily focusing on converting forecasting tasks into natural language processing tasks—the domain for which LLMs were originally designed and trained. Notably, Gruver et al. (2024) emphasized the importance of proper tokenization when employing LLMs for time series forecasting tasks. In their work, Gruver et al. highlighted that due to the fundamental mechanism of LLMs—predicting tokens, which are short segments of text defined uniquely by each model's tokenizer rather than full words—it is crucial to format input data appropriately before it is processed by the model. Their work specifically utilized a family of LLM known as Llama (Touvron et al., 2023), which employs a tokenizer significantly different from the tokenizer used by our LLM of interest, the GPT family (Radford and Narasimhan, 2018). Therefore, in our work, token considerations will play an essential role in our methodology, detailed in subsequent sections.

More recent research has focused on effectively incorporating contextual information into time series forecasting. A notable contribution in this direction is TimeLLM (Jin et al., 2023), which introduced the "Prompt-as-Prefix" method, where contextual information is appended prior to the sequence of forecast data, demonstrating significant effectiveness in improving the forecasting accuracy. Furthermore, Context-is-Key (CiK) (Williams et al., 2025) is a comprehensive benchmark that systematically evaluated and compared various frameworks that employ LLMs for time series forecasting, both those that incorporate prediction context and those that exclude it. Their primary finding indicated that LLM forecasting frameworks that incorporate prediction context substantially outperform those without contextual information. CiK also proposed a straightforward method termed "direct-prompt," where context and data are directly included in the prompts given to pre-trained LLMs. We intend to adopt a similar direct-prompting strategy as one of our LLM-based approach in this project.

2.3 Retrieval-Augmented Generation (RAG): Theory and Applications

Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) is a methodology that enhances the performance of Large Language Models (LLMs) by integrating external information sources, such as search results or relevant documents. When combined with pre-trained

LLMs, RAG has shown remarkable improvements in model performance not only for Natural Language Processing tasks but also for math-related tasks (Gao et al., 2024).

Interest in using RAG for time series forecasting has surged significantly over the past year. Notable works, such as the Retrieval Augmented Forecasting (RAF) framework (Tire et al., 2024), have introduced a well-defined and structured approach for RAG-enhanced LLM Time Series forecasting. RAF incorporates Amazon's time-series-specific LLM, Chronos (Ansari et al., 2024), and achieves a higher accuracy compared to the baseline Chronos. Our RAG-based forecasting framework and algorithms, particularly the rationale for selecting distance functions for retrieval-augmented generation, will be comparable to previous works while simultaneously incorporating novel approaches, as outlined in the subsequent chapters.

Zero-Shot Forecasting in a Retrieval-Augmented Generation Framework

Although Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) enriches the prompt with external context, it does not involve any adjustments or updates to the model's parameters. In our RAG-based forecasting framework, we retrieve relevant time series examples or domain knowledge at inference time and append them directly to the input prompt. Consequently, even with RAG, our forecasting remains a zero-shot application, only leveraging out-of-the-box LLMs to produce accurate forecasts.

CHAPTER 3

Data

3.1 Data Processing and Characteristics

Our study employs five publicly available CGM datasets that encompass a diverse range of diabetic and non-diabetic individuals. To ensure consistency and clarity, we refer to each dataset in text by the last name of the first author associated with the corresponding publication, followed by the publication year. In tables, we only include the last name of the first author. These datasets include Broll2021 (Broll et al., 2021), Colas2019 (Colás et al., 2019), Dubosson2018 (Dubosson et al., 2018), Hall2018 (Hall et al., 2018), and Weinstock2016 (Weinstock et al., 2016).

A summary of demographic statistics from the raw data of each dataset is shown in Table 3.1. It is worth noting that data availability and completeness vary across datasets. In particular, the Broll2021 and Dubosson2018 datasets lack certain metadata such as average age and sex distribution.

In our dataset selection, datasets featuring individuals with Type 1 diabetes are more prevalent than those featuring individuals with Type 2 diabetes. This is primarily due to the limited availability of CGM data for individuals with Type 2 diabetes, as CGMs are only recommended for Type 2 patients who are on insulin therapy. For datasets such as Colas2019 and Hall2018, the population is labeled as "mixed" because they include both

Table 3.1: Demographic information for each dataset

Dataset	**Diabetes**	**# of Subjects**	**Average Age**	**Sex (M/F)**
Broll	Type 2	5	NA	NA
Colas	Mixed	207	59	103/104
Dubosson	Type 1	9	NA	6/3
Hall	Mixed	57	48	25/32
Weinstock	Type 1	200	68	106/94

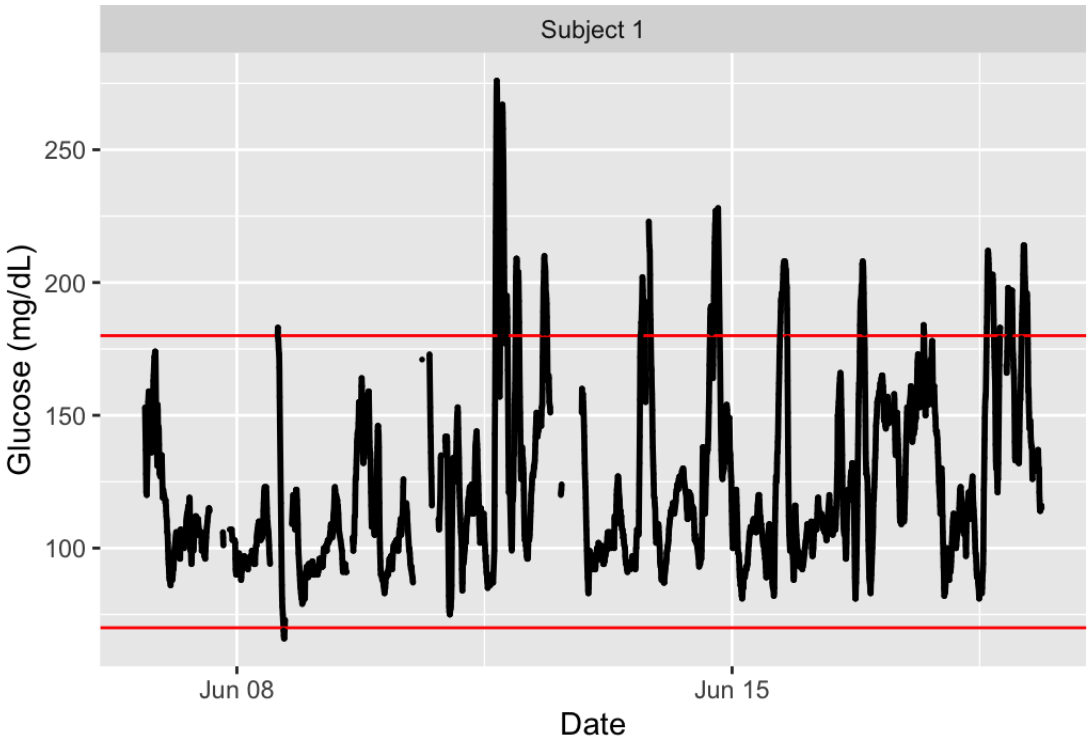


Figure 3.1: CGM readings for Subject 1 in Broll2021 dataset

healthy and prediabetic individuals, some of whom later progressed to Type 2 diabetes.

To facilitate initial exploration, we performed exploratory data analysis using the R package iglu (Broll et al., 2021). Figure 3.1 shows an example of a CGM time series plot from an individual in the Broll2021 dataset.

Several distinct characteristics can be observed from the provided time series plot. Firstly, the CGM readings exhibit pronounced fluctuations, rendering them inherently noisy. Secondly, segments of missing data are quite prevalent. Similar characteristics, particularly the high prevalence of missing data, appear to occur frequently across various datasets in our study.

3.2 Data Preprocessing

3.2.1 Interpolation & Segmentation

As previously mentioned, CGM datasets often exhibit multiple segments of missing data. This occurs when devices occasionally fail to record glucose readings due to sensor errors or user non-compliance, resulting in gaps in the time series. To address short-duration gaps,

Table 3.2: Interpolation and segmentation thresholds for each dataset

Dataset	**Broll**	**Colas**	**Dubosson**	**Hall**	**Weinstock**
Gap threshold (minutes)	45	45	30	30	45
Minimum segment length (hours)	20	16	20	16	20

we employ linear interpolation, which estimates missing values based on the glucose levels immediately preceding and following the gap.

Let each raw CGM segment be denoted

$$x^{(i)}_{j,1:L_j^{(i)}} \in \mathbb{R}^{L_j^{(i)}},$$

Where i is the index of the subjects, j is the index of continuous segments, and $L_j^{(i)}$ is the length of segment j. Suppose a missing gap of length $m+1$ occurs at indices k through $k+m$, i.e. the sub-slice

$$x^{(i)}_{j,k:k+m}$$

is unobserved, and that the adjacent values $x^{(i)}_{j,k-1}$ and $x^{(i)}_{j,k+m+1}$ are available. Denote the time duration of this gap by Δt.

Interpolation. If $\Delta t < \delta$ (the dataset-specific threshold), we impute each missing entry by linear interpolation:

$$x^{(i)}_{j,\,k+k'} = x^{(i)}_{j,\,k-1} + \frac{k'+1}{m+2}\left(x^{(i)}_{j,\,k+m+1} - x^{(i)}_{j,\,k-1}\right), \quad k' = 0, 1, \ldots, m.$$

Segmentation. If $\Delta t \geq \delta$, we consider the gap to be too large, making interpolation unreliable. In such cases, we divide the segment into two valid sub-segments:

$$\text{Segment A: } x^{(i)}_{j,\,1:k-1},$$
$$\text{Segment B: } x^{(i)}_{j,\,k+m+1:L_j^{(i)}}.$$

Each resulting segment is retained only if its length exceeds the dataset-specific minimum $L_{\min}$. This guarantees that all training and evaluation samples consist of sufficiently long, continuous CGM sequences. Following the guidelines provided by GlucoBench (Sergazinov et al., 2024), we define distinct interpolation thresholds and minimum segment durations for each dataset, as summarized in Table 3.2. Please note that readings are taken every five minutes. Consequently, a 45-minute gap threshold corresponds to nine consecutive missing readings, while a 30-minute gap threshold corresponds to six consecutive missing readings.

3.2.2 Data Splitting

To evaluate the accuracy of forecasting models, each dataset $\mathcal{D}$ is partitioned into four subsets: training $\mathcal{D}_{\text{tr}}$, validation $\mathcal{D}_{\text{val}}$, in-distribution test $\mathcal{D}_{\text{ID}}$, and out-of-distribution $\mathcal{D}_{\text{OD}}$. Ninety percent of subjects are randomly assigned to the in-distribution cohort, while the remaining ten percent form $\mathcal{D}_{\text{OD}}$. In this project, we exclude $\mathcal{D}_{\text{OD}}$ from analysis because our context-based LLM framework is prompt-based and cannot be used for out-of-distribution forecasting.

For each subject in the in-distribution cohort, CGM readings are ordered chronologically. The final 16 hours (192 measurements) are reserved for $\mathcal{D}_{\text{ID}}$, the preceding 16 hours for $\mathcal{D}_{\text{val}}$, and all earlier readings comprise $\mathcal{D}_{\text{tr}}$.

Thus, the full dataset is represented as

$$\mathcal{D} \;=\; \mathcal{D}_{\text{tr}} \;\cup\; \mathcal{D}_{\text{val}} \;\cup\; \mathcal{D}_{\text{ID}} \;\cup\; \mathcal{D}_{\text{OD}},$$

with $\mathcal{D}_{\text{OD}}$ held aside for future out-of-distribution evaluations.

CHAPTER 4

Method

4.1 Problem Definition

We formally define the Continuous Glucose Monitoring (CGM) data forecasting problem as follows. Let $\{x_j^{(i)}\}_{i,j}$ denote a sequence of CGM readings, with i indexing the subjects and j indexing the continuous segments. Then, Let $x_{j,k:k+L}^{(i)}$ be a length-L contiguous slice of a segment from index k to $k+L$. Our forecasting target is the subsequent T readings of the same segment, $y_{j,k+L+1:k+L+T}^{(i)}$, conditioned on the historical readings $x_{j,k:k+L}^{(i)}$, where T denotes the prediction horizon. Note that for a single segment $x_j^{(i)}$, we have $k+L+T \leq L_{min}$, where L_{min} is our pre-defined minimum value for a segment, which is unique to each dataset (Table 3.2).

Rather than relying on the entire segment of historical CGM readings, we treat the length of input window L as a hyperparameter to control how much recent history the model considers. This leads to a general formulation for the forecasting model $\mathcal{F}$, parameterized by θ, as follows:

$$\big(\hat{y}_{j,k+L+1}^{(i)}, \ldots, \hat{y}_{j,k+L+T}^{(i)}\big) \;=\; \mathcal{F}\big(x_{j,k}^{(i)}, \ldots, x_{j,k+L}^{(i)} \;;\; \theta\big).$$

Our choices of T and L are largely aligned with GlucoBench (Sergazinov et al., 2024). We want to forecast 1 hour ahead of the historical data, and since the time reading interval is 5 minutes, we select $T = 12$ for the entire project. As a hyperparameter, L is selected to optimize the performance of each forecasting model. Consequently, its value depends on the chosen forecasting model, as different models have varying capabilities in capturing patterns over extended periods of time series data.

4.2 Baseline Models

Four baseline, non-Large Language Model (LLM)-based forecasting models are chosen from the most successful models in GlucoBench's benchmark. These models comprise two shallow

methods: linear regression and Autoregressive Integrated Moving Average (ARIMA) (Box and Jenkins, 1976), as well as one deep learning approach, Transformer (Vaswani et al., 2023). Furthermore, a hybrid model (Latent ODE (Rubanova et al., 2019)) is included.

Training of the baseline models are performed on slices drawn from the training split $\mathcal{D}_{\text{tr}}$ and validated on the validation split $\mathcal{D}_{\text{val}}$. Concretely, for each subject i and a segment j in $\mathcal{D}_{\text{tr}}$, let

$$x^{(i)}_{j,1:L^{(i)}_j}, \qquad L^{(i)}_j > L_{\min},$$

denote the preprocessed CGM series. We uniformly sample starting indices

$$k \in \{1, 2, \ldots, L^{(i)}_j - L - T\}$$

without replacement. Each k yields an input window of length L,

$$x^{(i)}_{j,k:k+L} \;=\; \big(x^{(i)}_{j,k},\, x^{(i)}_{j,k+1},\, \ldots,\, x^{(i)}_{j,k+L}\big),$$

and the corresponding target of length T (with $T = 12$),

$$y^{(i)}_{j,k+L:k+L+T} \;=\; \big(y^{(i)}_{j,k+L+1},\, y^{(i)}_{j,k+L+2},\, \ldots,\, y^{(i)}_{j,k+L+T}\big).$$

A single input-output pair is then

$$\mathcal{P}_{\text{tr}} = \big(x^{(i)}_{j,k:k+L},\, y^{(i)}_{j,k+L+1:k+L+T}\big),$$

for each subject i and segment j in the training subset.

Finally, we combine all the training input-output pairs and train the forecasting model on this aggregated training set. This model is then validated on $\mathcal{D}_{\text{val}}$ and evaluated on $\mathcal{D}_{\text{ID}}$. The methodology used for model evaluation on $\mathcal{D}_{\text{ID}}$ is detailed in the following section.

4.3 Direct-Prompt Forecasting Framework for Large Language Models

In this section, I formalise a direct-prompt strategy akin to the one described in Williams et al. (2025) that employs a pre-trained LLM to forecast the next observation of a CGM time series sequence. The procedure is conceptually simple—one provides natural-language prompt containing the most recent L CGM readings, preceded by relevant contextual information, and asks the model to produce the subsequent value.

First, a sequence of CGM readings started at index k $x_k, \ldots, x_{k+L} \in \mathbb{R}$ is first converted to comma separated numbers (e.g. "127, "), then passed through the LLM's tokenizer. This produces the sequence $\mathbf{r} = (r_1, \ldots, r_n)$, a tokenised sequence of length n that encodes the historical CGM readings. Note that n need not equal L, since LLMs will potentially split a single numeric value into multiple tokens, especially when numbers or formatting (such as commas) differ significantly from tokens present in the tokenizer's vocabulary.

Then,

$$\mathbf{c} = \big(c_1, \ldots, c_m\big), \qquad \mathbf{r} = \big(r_1, \ldots, r_n\big),$$

are two disjoint token sequences, where $\mathbf{c}$ encodes contextual and demographic information (e.g. diabetes type, age, lifestyle factors) and $\mathbf{r}$ encodes the numeric CGM readings after tokenisation. Here, m is the number of context tokens and n is the number of tokens covering those L readings. We then concatenate these to form the single prompt

$$\mathbf{s} = \big(c_1, \ldots, c_m,\ r_1, \ldots, r_n\big).$$

A pre-trained LLM with parameters θ has been trained (during pre-training phase) to maximize the autoregressive log-likelihood of sequences:

$$\mathcal{L}(\theta) = \sum_{t=1}^{m+n} \log p_\theta\big(s_t \,|\, s_{<t}\big),$$

where $s_{<t} = (s_1, \ldots, s_{t-1})$. Here the index t runs over all tokens in the prompt: the first m positions correspond to context tokens, the next n to historical readings.

At inference, we condition on the observed prompt $\mathbf{s}$ and generate the next T' tokens

$$s_{m+n+1},\ s_{m+n+2},\ \ldots,\ s_{m+n+T'},$$

by sampling from $p_\theta(s_t \mid s_{<t})$ for $t > m + n$. Again, T' need not equal the forecasting horizon T. Each generated token $s_{m+n+j'}$ is then mapped back to a numeric glucose value $\hat{y}_j$ via the inverse of the tokenization scheme. Consequently, we obtain the forecasted sequence

$$\big(\hat{y}_{k+L+1}, \hat{y}_{k+L+2}, \ldots, \hat{y}_{k+L+T}\big),$$

which corresponds to the next T CGM measurements following the input window. Note that during inference, no tuning of parameters or training occur, hence, we will not be using $\mathcal{D}_{\text{tr}}$ in the direct-prompt framework. Instead, we directly evaluate the model on $\mathcal{D}_{\text{ID}}$.

An illustration of the Direct-Prompt Framework is shown in 4.1.

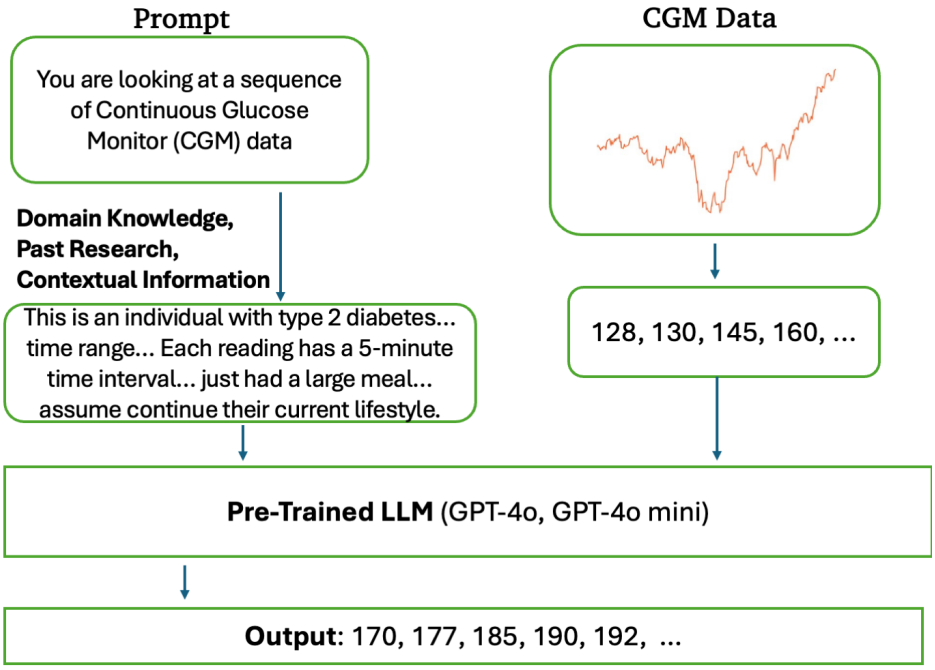


Figure 4.1: Illustration of the Direct-Prompt Framework

LLM Selection and Configuration

For the direct-prompt framework, we employ two pre-trained models from the GPT family: GPT-4o and GPT-4o-mini (Radford and Narasimhan, 2018). Both models share the same tokenizer and decoder architecture, differing only in their total parameter count. This choice allows us to examine whether forecasting performance of LLMs systematically vary with their model size.

To capture time-sensitive patterns in CGM data, where glucose levels frequently exhibit periodicity over a single day, we set the input window length to $L = 144$ tokens (which corresponds to 12 hours with five-minute intervals). All hyperparameters that control the "creativity" of LLM, such as temperature, top-k, and top-p, will remain at their default values.

4.3.1 Contextual Prompt Engineering for Forecasting

As the direct-prompt framework is a zero-shot forecasting method that does not necessitate any additional tuning or training, a meticulously crafted prompt that incorporates both numerical history and all contextual cues becomes essential for accurate forecasting. Drawing inspiration from the prompt design proposed in Time-LLM(Jin et al., 2023), we structure each prompt into four distinct sections:

1. **Contextual information** We explicitly state key demographics—diabetes type, age, sex, insulin-therapy status—so that the LLM can condition its prior on patient-specific physiology.

2. **Physiological constraints** The model is reminded that glucose values are strictly positive and constrained by device limits (CGM devices typically have a measuring range between 40–400mg/dL). This reduces out-of-range hallucinations.

3. **Scenario descriptors (optional)** When available, we include scenario factors such as a large meal, vigorous exercise, or an insulin injection, enabling the model to anticipate short-term deviations.

4. **Output instruction** A direct instruction on output format (e.g. "Return three numbers separated by commas; no commentary.") leads LLM to provide a parseable, numeric-only response.

Below, we illustrate an example prompt from the Broll2021 Dataset.

Example from the Broll2021 Dataset

You are looking at a sequence of Continuous Glucose Monitor (CGM) data. This is an individual with type 2 diabetes. Each reading has a 5-minute time interval, and the first reading in the given sequence is recorded at [*timestamp*].
Glucose levels are positive, with an upper measuring limit of 400 mg/dL and a lower limit of 40 mg/dL. Assume the patient continues with their current lifestyle.
Predict the next 12 readings. Return the forecast in just numbers. Do not include any other information(e.g., comments) in the forecast.

Historical Readings: 128, 130, 145, 160, ...

4.4 Retrieval-Augmented Generation Forecasting

Building on the direct-prompt paradigm, we introduce a Retrieval-Augmented Generation (RAG) framework (Lewis et al., 2021) that enriches the prompt with sampled historical CGM trajectories. Retaining the core mechanics of the direct-prompt approach, our RAG pipeline employs a distance-metric-based retriever function to retrieve and embed time-series examples that closely mirror the current CGM window, thereby supplying the LLM with analogues of the sequence it must forecast.

Let a tokenised CGM segment for subject i, segment j in the training dataset $\mathcal{D}_{\text{tr}}$ be

$$x^{(i)}_{j,1:L^{(i)}_j}, \quad L^{(i)}_j > L_{\min}.$$

Using a sliding window of length L and stride of s, we extract input–output pairs of subject i from every segment j in $\mathcal{D}_{\text{tr}}$:

$$\big(x^{(i)}_{j,k:k+L},\, y^{(i)}_{j,k+L:k+L+T}\big), \quad k \in \{\,1,\, 1+s,\, 1+2s, \ldots,\, L^{(i)}_j - L - T\}\,,$$

where $y^{(i)}_{j,k+L:k+L+T}$ contains the next T readings ($T = 12$). Treat each input slice $(x^{(i)}_{j,k:k+L})$ as a key $\mathbf{k}_\ell$ and the concatenation $(x^{(i)}_{j,k:k+L}, y^{(i)}_{j,k+L:k+L+T})$ as its value $\mathbf{v}_\ell$. Collecting all such pairs yields a vector database of past CGM readings for subject i

$$\mathcal{M}_i \;=\; \{\,(\mathbf{k}_\ell, \mathbf{v}_\ell)\,\}_{\ell=1}^{M}.$$

Query and Retrieval

Let $\mathbf{x}_{\text{cur}} = \big(x^{(\star)}_k, \ldots, x^{(\star)}_{k+L}\big)$ be the CGM reading sequence for an arbitrary test instance starting at index k of input length L that we want to forecast, so that $\mathbf{x}_{\text{cur}}$ and $\mathbf{k}_\ell$ both in $\mathbb{R}^L$. We can construct a retriever by ranking all keys using a specified distance metric and retrieving the K closest neighbors $\mathcal{V}_K$:

$$\mathcal{V}_K(\mathbf{x}_{\text{cur}}) = \arg \min_{\substack{I \subseteq \{1,\ldots,M\} \\ |I| = K}} \sum_{\ell \in I} d\big(\mathbf{x}_{\text{cur}}, \mathbf{k}_\ell\big).$$

where d is (i) Euclidean distance, (ii) Mean Absolute Percentage Error(MAPE), or (iii) 1−Pearson correlation, defined respectively as

$$d_{\text{Euclid}}(\mathbf{x}_{\text{cur}}, \mathbf{k}_\ell) = \big\|\mathbf{x}_{\text{cur}} - \mathbf{k}_\ell\big\|_2 = \sqrt{\sum_{t=1}^{L} \big(x_{\text{cur},t} - k_{\ell,t}\big)^2}.$$

$$d_{\text{MAPE}}(\mathbf{x}_{\text{cur}}, \mathbf{k}_\ell) = \frac{1}{L} \sum_{t=1}^{L} \left| \frac{x_{\text{cur},t} - k_{\ell,t}}{x_{\text{cur},t}} \right| \quad \big(x_{\text{cur},t} \neq 0\big).$$

$$d_{\text{corr}}(\mathbf{x}_{\text{cur}}, \mathbf{k}_\ell) = 1 - \rho\big(\mathbf{x}_{\text{cur}}, \mathbf{k}_\ell\big) = 1 - \frac{\sum_{t=1}^{L} \big(x_{\text{cur},t} - \bar{x}_{\text{cur}}\big)\big(k_{\ell,t} - \bar{k}_\ell\big)}{\sqrt{\sum_{t=1}^{L} \big(x_{\text{cur},t} - \bar{x}_{\text{cur}}\big)^2}\, \sqrt{\sum_{t=1}^{L} \big(k_{\ell,t} - \bar{k}_\ell\big)^2}}.$$

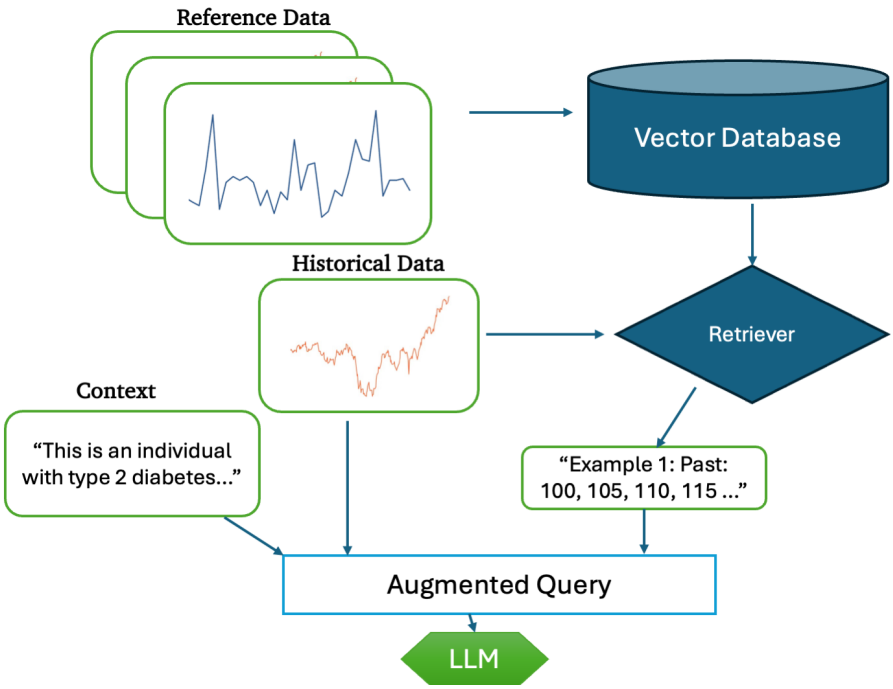


Figure 4.2: Illustration of the RAG approach

Augmented Prompt and Generation

We then prepend the retrieved examples to the existing prompt. In addition to the disjoint token sequences $\mathbf{c} = (c_1, \ldots, c_m)$, which encodes the context, and $\mathbf{r} = (r_1, \ldots, r_n)$, which encodes the historical CGM readings after tokenisation, as defined in the direct-prompt framework, we retrieve K past sequences $\mathcal{V}_K = \{\mathbf{v}_\ell\}_{\ell=1}^{K}$, where each sequence is tokenized as

$$\mathbf{v}_\ell = (v_{\ell,1}, \ldots, v_{\ell,n}), \qquad \ell = 1, \ldots, K.$$

These retrieved sequences are inserted immediately before the histoical CGM reading tokens $\mathbf{r}$, yielding the complete prompt $\mathbf{s}_{\text{RAG}}$ for our RAG framework:

$$\mathbf{s}_{\text{RAG}} = \big(c_1, \ldots, c_m,\ v_{1,1}, \ldots, v_{1,n},\ \ldots,\ v_{K,1}, \ldots, v_{K,n},\ r_1, \ldots, r_n\big).$$

As in the direct-prompt framework, the model then generates the forecasted sequence conditioned on $\mathbf{s}_{\text{RAG}}$.

Few-Shot Prompting

To incorporate the retrieved historical data into our prediction, we employ a method known as few-shot prompting. In few-shot prompting, the retrieved segments serve as examples

demonstrating the forecasting pattern the LLM should adopt. This method has been demonstrated to enhance the performance of LLMs in prediction and classification tasks (Dong et al., 2024). The retrieved pairs serve as in-context exemplars, transforming forecasting into a Pattern learning problem for LLMs. For instance, a simplified prompt using an example of historical length $L = 4$ and forecast horizon $T = 3$ is illustrated below:

RAG Prompt Example

You are looking at a sequence of Continuous Glucose Monitoring (CGM) data...

...

Predict the next 3 readings. Return the forecast in just numbers. Do not include any other information (e.g., comments) in the forecast.

Example 1 — Past: 100, 105, 110, 115 → Future: 120, 125, 130

Example 2 — Past: 104, 108, 113, 117 → Future: 122, 127, 132

Current readings: — Past: 102, 107, 112, 118 → Future: *[Forecast]*

4.4.1 Selection of Hyperparameters

In our retrieval experiments, we select $K = 3$ exemplars per query. Empirically, retrieving fewer than three examples yields insufficient information, while larger K values often lead to overfitting and decreased accuracy. For the vector index, we employ a sliding-window stride of $s = 6$ readings (30 minutes), as smaller strides result in highly overlapping segments that provide minimal additional information.

We also examine two input window lengths, L, to gauge the sensitivity of distance-based retrieval to series length. First, $L = 144$ tokens (12 hours at 5-minute intervals) matches the direct-prompt baseline. Second, $L = 72$ tokens (6 hours) focuses retrieval on shorter, potentially more distinctive CGM patterns.

Similar to the direct prompt framework, we retained all hyperparameters (e.g., temperature, top-k, top-p) at their default values.

4.5 Evaluation of Model Performance

Model accuracy is measured on the in-distribution test set $\mathcal{D}_{\text{ID}}$ with the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE):

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{j=1}^{T}\left(y_{k+L+j} - \hat{y}_{k+L+j}\right)^2}, \qquad \text{MAE} = \frac{1}{T}\sum_{j=1}^{T}|y_{k+L+j} - \hat{y}_{k+L+j}|,$$

where T is the prediction horizon. For each subject, we compute RMSE and MAE, then report the median across subjects to mitigate the effect of skewed error distributions. This procedure is applied to both baseline models and LLM-based frameworks.

Evaluation of LLM-based methods

Because LLM outputs are nondeterministic and stochastic in nature (Song et al., 2024), we generate five independent forecasts for every segment, average the predictions point-wise,

$$\bar{y}_t = \frac{1}{5} \sum_{i=1}^{5} \hat{y}_t,$$

and compute errors with averaged prediction $\bar{y}_t$ instead of $\hat{y}_t$ to reduce the variance in prediction.

To gain a comprehensive understanding of the contextual comprehension abilities of LLMs, we conduct an analysis of randomly selected individual forecasts under specific prompts. In addition to real-world scenarios, we introduce two synthetic scenarios: "patient has just had a big meal" and "patient has just received an insulin injection." By comparing the models' responses to physiological expectations, we evaluate their ability to integrate prompt information and adapt to various scenarios.

CHAPTER 5

Results

5.1 Individual Forecasting Results

5.1.1 Context-Based Forecasting Outcomes

We evaluate whether LLM accurately incorporated the demographic cues presented in the prompt. To demonstrate the impact of contextual information, we randomly selected a 24-hour segment from a subject in the Dubosson2018 dataset. It's important to note that all participants in this dataset have type-1 diabetes, and the chosen subject is known to be on insulin therapy. While the figures include 24 hours of the segment, only $L = 144$, or 12 hours of readings, are used as input length.

Baseline prompt (insulin status not mentioned)

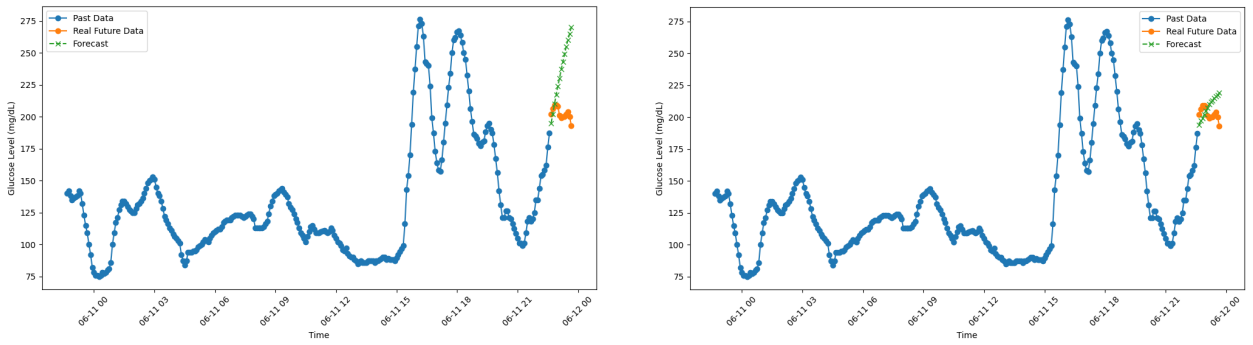
Figure 5.1a shows the forecast obtained when the prompt contains no reference to insulin therapy.

The model largely forecasts the recent upward trend, predicting a rise toward the previous maximum. It fails, however, to anticipate the sharp drop occurring roughly one hour ahead—an unsurprising limitation given the high variability of CGM data, especially for diabetic subjects.

Context-aware prompt (insulin status specified).

We then repeated the forecast with an identical prompt except for a single sentence stating that the subject is currently receiving insulin therapy.

The LLM immediately moderates its prediction: the projected slope is less steep and aligns more closely with the ground-truth trajectory. This qualitative improvement confirms that the model does attend to, and act upon, the demographic context embedded in the prompt.



(a) Baseline (no insulin)

(b) Insulin status specified

Figure 5.1: LLM forecasts for the same Dubosson2018 segment (12 h input) under (a) a baseline prompt and (b) an insulin-aware prompt.

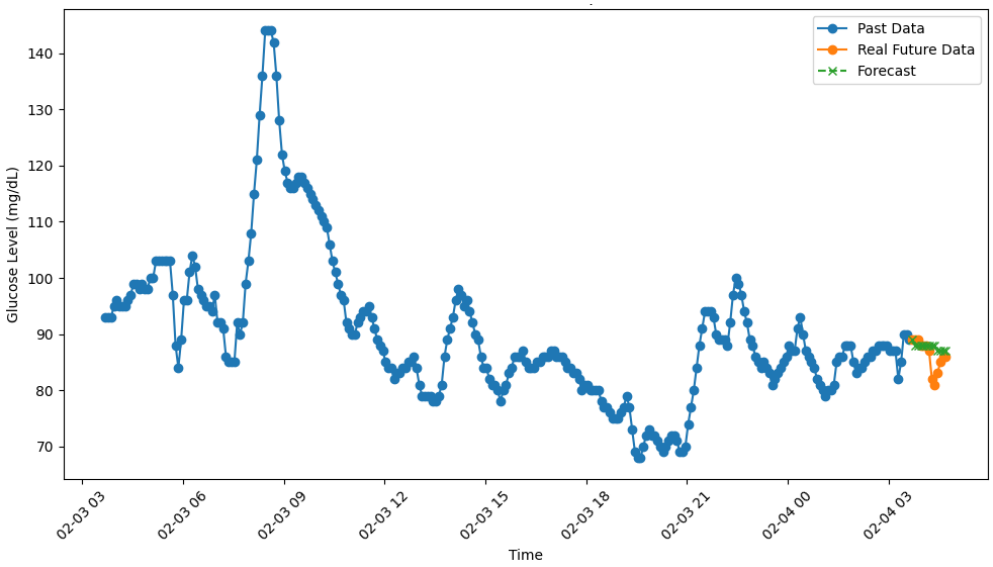


Figure 5.2: LLM forecast for a healthy subject

Healthy-subject

For comparison, we present a forecast for a healthy individual drawn from the Hall2018 dataset in Figure 5.2.

When the prompt specifies a healthy, non-diabetic subject, the LLM adapts to this context by expecting smoother glucose dynamics and therefore generates forecasts that closely follow the true trajectory.

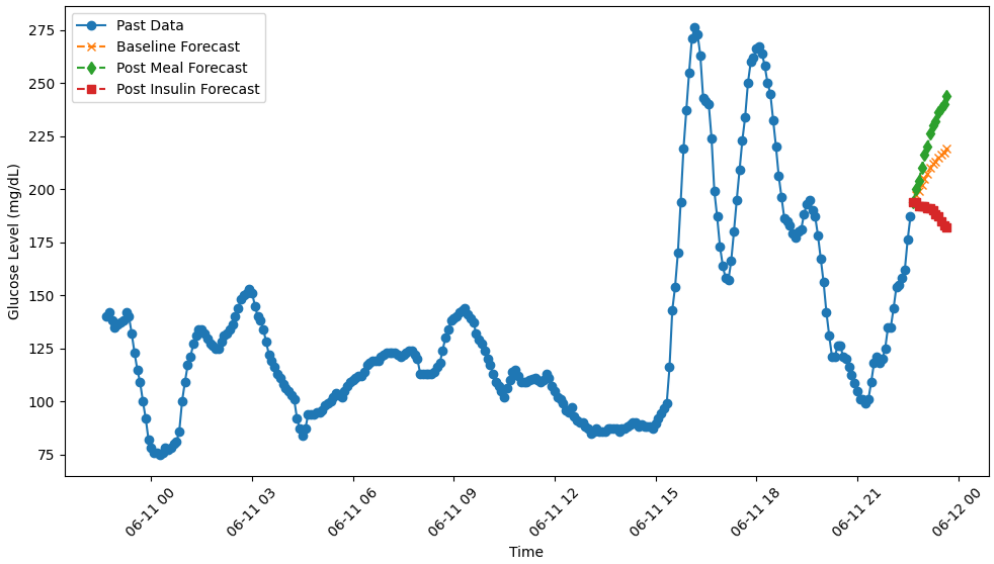


Figure 5.3: Scenario-based forecasts for the Dubosson2018 subject: normal, post-meal, and post-insulin injection.

5.1.2 Scenario-Based Forecasting Outcomes

Finally, we explored the model's capacity to reason over hypothetical scenarios. Using the same Dubosson2018 segment we have randomly chosen in previous section, we asked LLM to generate three forecasts by modifying the prompt: (i) normal conditions, (ii) the subject has just consumed a large meal, and (iii) the subject has just administered a dose of insulin.

The scenario-based forecasts in Figure 5.3 show that the LLM adapts its predictions to each condition: when prompted "after the meal," it projects a pronounced glucose spike; under "normal" conditions, it forecasts a moderate rise; and following "after insulin injection," it anticipates a clear downward trend. These results align with the expected physiological responses and demonstrate that LLM is capable of accurately modifying forecasts based on real-time scenarios.

5.2 Aggregate Dataset Results

5.2.1 Direct-Prompt Forecasting

Table 5.1 summarizes the predictive performance of four classical models (ARIMA, linear regression, Transformer, Latent ODE) and two direct-prompt GPT-4o variants (GPT-4o, GPT-4o-mini). For each dataset, the lowest root mean squared error (RMSE) and mean absolute error (MAE) are bolded.

Table 5.1: Prediction accuracy of baseline and direct-prompt models across datasets

Accuracy	**Colas**		**Dubosson**		**Hall**		**Broll**		**Weinstock**	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
ARIMA	5.81	4.88	13.16	11.30	8.68	7.35	12.18	10.29	14.35	11.09
Linear	**5.28**	**4.36**	**12.20**	**8.68**	7.34	6.30	12.20	9.91	13.71	11.61
Transformer	6.29	5.44	16.75	13.17	7.54	6.44	13.32	11.66	**13.23**	**11.24**
Latent ODE	6.46	5.65	18.75	16.62	7.15	6.12	14.23	12.47	13.64	11.45
GPT-4o	9.20	8.06	27.27	24.40	8.18	7.11	**4.71**	**3.97**	16.25	13.33
GPT-4o-mini	5.95	5.08	19.26	15.03	**6.45**	**5.32**	8.96	6.88	13.25	11.57

Table 5.2: Best RAG configurations and respective prediction accuracy

Accuracy	**Colas**		**Dubosson**		**Hall**		**Broll**		**Weinstock**	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
4o, 72, MAPE	**7.18**	**5.92**	15.73	14.83	5.85	5.07	16.32	13.64	15.51	12.82
4o, 144, MAPE	7.79	6.64	**12.23**	**9.40**	7.49	6.50	4.93	4.03	17.91	14.64
4o, 72, Euclid.	7.72	6.89	18.59	15.66	**5.48**	**4.54**	8.90	7.32	13.95	12.09
4o, 144, Corr.	8.10	7.06	22.32	18.92	6.85	6.03	**4.71**	**3.97**	16.25	13.33
4o, 72, Corr.	7.19	6.22	15.89	14.18	6.21	5.17	10.19	8.75	**13.72**	**12.16**

The simple linear regression is the best-performing model for Colas2019 and Dubosson2018 datasets. On Hall2018 (mixed dataset including normoglycemic and prediabetic individuals), GPT-4o-mini has the highest accuracy. On Broll2021 (the smallest dataset including type-2 diabetic individuals), GPT-4o substantially improves the RMSE and MAE by a considerable margin, while GPT-4o-mini lags behind but still outperforms all the baseline models. The Transformer is the best-performing model on Weinstock2016 (the largest dataset), but GPT-4o-mini closely trails behind in terms of RMSE and MAE.

5.2.2 Retrieval-Augmented Generation

Best RAG configurations and their Prediction Accuracy

Table 5.2 lists the five Retrieval-Augmented Generation (RAG) configurations that have achieved the lowest RMSE and MAE in at least one dataset. The settings for each RAG configuration, represented as $(\text{model}, L, \text{metric})$, are presented in the first column. GPT-4o and GPT-4o-mini are abbreviated as 4o and 4o-mini, respectively. Similarly, pearson correlation, euclidean distance, and mean absolute percentage error (MAPE) are abbreviated as Corr., Eucli., and MAPE, respectively.

None of the RAG-enhanced GPT-4o-mini configurations are selected in this table, due

Table 5.3: GPT-4o (direct-prompt) vs. its RAG variants. RAG cells show percentage change relative to GPT-4o (blue = improvement, red = deterioration).

Change	**Colas**		**Dubosson**		**Hall**		**Broll**		**Weinstock**	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
GPT-4o	9.20	8.06	27.27	24.40	8.18	7.11	4.71	3.97	16.25	13.33
4o, 72, MAPE	-22%	-26%	-42%	-39%	-28%	-29%	+246%	+244%	-5%	-4%
4o, 144, MAPE	-15%	-18%	-55%	-61%	-8%	-9%	+5%	+2%	+10%	+10%
4o, 72, Euclid.	-16%	-15%	-32%	-36%	-33%	-36%	+89%	+84%	-14%	-9%
4o, 144, Corr.	-12%	-12%	-18%	-22%	-16%	-15%	0%	0%	0%	0%
4o, 72, Corr.	-22%	-23%	-42%	-42%	-24%	-27%	+116%	+120%	-16%	-9%

to the lack of performance of GPT-4o-mini under the RAG framework. All the following observations are specifically about RAG framework with GPT-4o.

We observe that the best-performing RAG configurations vary across datasets and strongly depend on the chosen distance metric and input window length. Specifically, MAPE achieves the strongest results for Colas2019 (with an input length of 72) and Dubosson2018 (input length of 144). Pearson correlation is most effective for Broll2021 (input length of 72) and Weinstock2016 (input length of 144). Meanwhile, Euclidean distance notably performs best on the Hall2018 dataset (input length of 72).

The comprehensive result of prediction accuracy for every RAG configuration is presented in the appendix A.

RAG vs. Direct-Prompt

Additionally, we quantify the improvements achieved by RAG configurations relative to GPT-4o's direct-prompt results. Specifically, we calculate the percentage change in RMSE and MAE for each RAG configuration compared to the baseline GPT-4o direct-prompt framework, providing a clear measure of the extent to which retrieval augmentation enhances forecasting accuracy.

Table 5.3 demonstrates that RAG produces substantial reductions in RMSE and MAE for Colas2019, Dubosson2018, and Hall2018. The effect is most pronounced in Dubosson2018, where RAG achieves reductions of up to 55% in RMSE and 61% in MAE. Weinstock2016 also benefits from RAG, though the improvements are more modest by comparison. In contrast, Broll2021 is the only dataset where RAG results in a notable increase in both RMSE and MAE. Notably, the GPT-4o configuration using pearson correlation with an input length of 72 delivers nearly identical performance to the direct-prompt baseline on both Broll2021 and Weinstock2016.

RAG vs. Baselines

Additionally, we compare the performance of RAG-enhanced models against the best-performing baseline models reported in GlucoBench (Sergazinov et al., 2024). For each dataset, we select the most effective RAG configuration (Table 5.3) and compute the percentage reduction in RMSE and MAE relative to the top-perfoming baseline method.

Table 5.4: Percentage changes (baseline vs. best RAG configurations) with coloring (blue = improvement, red = deterioration)

	Colas		Dubosson		Hall		Broll		Weinstock	
Model	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Best RAG configurations	4o, 72, MAPE		4o, 144, MAPE		4o, 72, Euclid.		4o, 144, Corr.		4o, 72, Corr.	
ARIMA	+23.6%	+21.3%	-7.1%	-16.8%	-36.9%	-38.2%	-61.4%	-61.4%	-4.4%	+9.0%
Linear	+36.0%	+35.8%	+0.2%	+8.3%	-25.4%	-27.9%	-61.4%	-60.0%	+0.1%	+4.1%
Transformer	+14.1%	+8.8%	-27.0%	-28.6%	-27.3%	-29.5%	-64.7%	-66.0%	+3.7%	+7.6%
Latent ODE	+11.2%	+4.8%	-34.8%	-43.5%	-23.4%	-25.8%	-66.9%	-68.2%	+0.6%	+5.6%

As shown in Table 5.4, the best-performing RAG configurations outperform baseline models on both Hall2018 and Broll2021, achieving substantial improvements—specifically, a 61% to 67% reduction in RMSE on each dataset. On Dubosson2018, RAG also yields meaningful reductions in RMSE and MAE relative to all baselines except linear regression. In contrast, Colas2019 and Weinstock2016 are the two datasets where RAG-based methods offer limited or marginal improvements.

5.2.3 Summary of Best-Performing Models

Table 5.5 shows the best forecasting models for each dataset, including baseline methods, direct-prompt, and RAG-enhanced LLMs. Linear regression tops Colas2019 and Dubosson2018, while the Transformer model leads on Weinstock2016. Two RAG configurations achieve the best results on Hall2018 and Broll2021.

Table 5.5: Best models or RAG configurations across datasets

Accuracy	**Colas**		**Dubosson**		**Hall**		**Broll**		**Weinstock**	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Linear	**5.28**	**4.36**	**12.20**	**8.68**	7.34	6.30	12.20	9.91	13.71	11.61
Transformer	6.29	5.44	16.75	13.17	7.54	6.44	13.32	11.66	**13.23**	**11.24**
4o, 144, Corr.	8.10	7.06	22.32	18.92	6.85	6.03	**4.71**	**3.97**	16.25	13.33
4o, 72, Euclid.	7.72	6.89	18.59	15.66	**5.48**	**4.54**	8.90	7.32	13.95	12.09

CHAPTER 6

Discussion

6.1 Strengths of the Direct-Prompt Framework

6.1.1 Dynamic Adaptation to Contextual Scenarios

Our assessment of individual forecasting results highlights a key advantage of Large Language Models (LLMs) in continuous glucose monitoring (CGM) forecasting: their ability to integrate and respond dynamically to contextual information without re-training. Unlike classical methods such as ARIMA or linear regression, which model glucose solely as a univariate stochastic process dependent on past measurements, LLMs inherently utilize demographic information, lifestyle factors, and device characteristics provided directly within the prompt. In practical terms, when an LLM receives a prompt such as "type 2 diabetes patient, post-meal state, device limit of 40-400 mg/dL," it leverages its extensive pre-trained knowledge to infer how these factors modulate glycemic variability. This contextual awareness is particularly valuable in CGM forecasting, as diabetes type, insulin therapy, and meal timing all introduce deviations in glucose trajectories. Additionally, scenario-based forecasting underscores another distinctive benefit of LLMs: their capability to immediately incorporate real-time events, such as "large meal" consumption or "insulin injection," into forecasts without the need for re-training. This is a distinctive feature of LLMs—one that is virtually absent from all baseline models. It resembles a Bayesian framework, in which prior knowledge is continuously updated with new, scenario-specific information about the individual—enabling the model to adapt its predictions in a context-sensitive manner, all without the need for re-training.

6.1.2 Accuracy of Direct-Prompt Forecasting

Impact of Model Size (GPT-4o vs. GPT-4o-mini)

The results across the five public CGM datasets indicate that GPT-4o-mini achieves strong zero-shot forecasting performance under the direct-prompt framework, consistently outperforming GPT-4o on all datasets except Broll2021. A plausible explanation for this outcome is that GPT-4o-mini, being a smaller LLM with fewer parameters, is inherently more calibrated and precise for numeric predictions in zero-shot contexts. In contrast, GPT-4o, with its substantially larger parameter count, tends to overly interpret subtle or ambiguous cues from the prompt, potentially incorporating irrelevant context or even "hallucinating" in its predictions. Consequently, within the direct-prompt approach, there emerges a clear advantage for smaller, more focused LLMs such as GPT-4o-mini, due to their reduced susceptibility to context-driven overfitting.

Dataset-Specific Effects on Forecast Performance

On the Broll2021 dataset, both GPT-4o and GPT-4o-mini under the direct-prompt framework achieve outstanding accuracy, surpassing all classical baselines by huge margins. As the smallest dataset in our study—comprised solely of type-2 diabetes subjects with highly variable glucose dynamics—Broll2021 showcases the power of LLMs' vast pre-trained knowledge and their ability to make strong forecasts from very limited historical observations. With the assistant of proper prompts, LLMs can extrapolate patterns that traditional models, which depend on extensive in-sample training, simply cannot capture. Notably, this is the only dataset where GPT-4o outperforms GPT-4o-mini, suggesting that its larger parameter capacity provides an advantage in reasoning over sparse and highly variable data, enabling more accurate zero-shot predictions in such settings.

Additionally, GPT-4o-mini outperforms all other models on the Hall2018 dataset, a moderately sized collection primarily consisting of healthy individuals. It also closely trails the best-performing model (Transformer) on Weinstock2016, the largest dataset analyzed. These results indicate that the predictive capabilities of LLMs extend beyond simple extrapolation; they demonstrate a capacity to recognize and utilize complex time series patterns and signals, typically captured only by deeper, explicitly trained models.

By contrast, on the Colas2019 and Dubosson2018 datasets, GPT-4o and GPT-4o-mini under the direct-prompt framework yield less accurate forecasts than simple linear regression. Notably, other complex models in the baseline suffer a similar fate, suggesting that the key challenge is not the absence of "prompt engineering," but the tendency of complex models to overfit noisy, low-signal time series. Even in zero-shot forecasting, LLMs can over-interpret

random fluctuations or spurious prompt cues, producing unstable predictions. In these datasets, the robustness of linear models outweighs the encyclopedic reasoning power of LLMs.

6.2 Evaluating Retrieval-Augmented Generation

Influence of Model Scale and Configuration

When incorporating RAG with GPT-4o and GPT-4o-mini, model size once again emerges as a crucial determinant of performance. GPT-4o, with its larger parameter capacity, leverages retrieved examples to reduce prediction errors across nearly all datasets and distance metrics. Conversely, GPT-4o-mini frequently struggles to extract meaningful signals from retrieved historical data, occasionally resulting in decreased accuracy. Its smaller architecture limits its capacity to effectively reason over multiple historical windows or adapt to nuanced local trajectory patterns. This mirrors our findings from the direct-prompt approach, reinforcing the idea that adequately large models are essential for fully capitalizing on RAG's potential in CGM forecasting.

We also find that the choice of retrieval metric and input-window length substantially influences forecasting performance. MAPE and Pearson correlation consistently deliver the largest reductions in RMSE and MAE, likely because these metrics effectively capture proportional and shape-based similarities in glucose trajectories. Euclidean distance, by contrast, struggles on longer input length ($L = 144$); however, it occasionally surpasses the other two distance metrics, on shorter input length ($L = 72$), likely due to its effectiveness in capturing immediate, localized trends and similarities over brief intervals.

Dataset-Dependent Performance of RAG

Similar to what we observed in direct-prompt, RAG's benefits prove highly dataset-dependent. While GPT-4o and GPT-4o-mini achieve strong performance on Broll2021 under direct-prompting, their accuracy drops significantly when RAG is applied in this dataset. This decline can be attributed to the dataset's small size and high variability in the time series curve—characteristics that make it difficult to retrieve representative historical segments. As a result, the retrieved examples can often introduce misleading noise rather than meaningful context, ultimately degrading forecast accuracy.

By contrast, on all other datasets—even the relatively small Dubosson2018—RAG yields consistent and notable improvements. Once a dataset reaches a moderate size and includes sufficiently representative historical segments, RAG reliably reduces both RMSE and MAE.

This pattern suggests while RAG can significantly enhance zero-shot forecasting, its effectiveness depends on the availability of a large, stable pool of past examples that provide relevant and informative context.

Notably, a RAG configuration achieves nearly identical performance to the direct-prompt baseline on both Broll2021 and Weinstock2016—the two datasets where RAG failed to improve, or only marginally improved, forecasting accuracy. While the underlying reason remains unclear, we hypothesize that the LLM may learn to disregard retrieved segments when it deems them unhelpful for prediction. This suggests a potential internal mechanism for selectively filtering irrelevant context. Further research is needed to better understand how LLMs weigh retrieved information during inference and under what conditions they choose to rely on or ignore it.

6.3 Model Recommendations and Practical Guidance

For broad-scale CGM forecasting with minimal engineering overhead, we recommend GPT-4o-mini in a direct-prompt setup. While it may not be the top-performing model on every dataset, it consistently delivers reliable zero-shot forecasts that are comparable to—or even better than—those of traditional baseline models. Importantly, it does so at a significantly lower cost than RAG configurations involving GPT-4o. Furthermore, as a zero-shot approach, it requires no training or fine-tuning, offering a major advantage over deep learning models like Transformers, which demand substantial computational resources and setup time.

GPT-4o-based RAG should be strategically reserved for enhancing predictive accuracy on specific datasets. Through additional fine-tuning and cross-validation, RAG has the potential to attain or even surpass the accuracy of the most performing baseline models.

Finally, in datasets where deep models frequently overfit, such as Colas2019 and Dubosson2018, simple linear regression continues to be the most reliable option.

6.4 Cost-Efficient and Clinically Scalable Forecasting with LLMs

While predictive accuracy is essential, the practicality of deploying forecasting models in real-world clinical or consumer settings is equally important. LLMs—whether operating under a direct-prompt or RAG framework—offer substantial advantages in this regard. Unlike traditional forecasting pipelines, which demand significant computational resources for model

training, hyperparameter tuning, and ongoing maintenance, zero-shot LLM forecasts can be generated with a single API call.

In our experiments, each forecast involved approximately 200 input tokens and 30 output tokens. At current GPT-4o pricing, this setup costs roughly $0.01 per forecast, with GPT-4o-mini offering even greater affordability.

From a clinical standpoint, LLMs' ability to incorporate rich contextual information into forecasts enables dynamic, patient-specific glucose predictions. Furthermore, their deployment via secure, cloud-based APIs allows existing CGM platforms to stream real-time glucose data and contextual cues directly to an LLM endpoint. These low-latency, low-cost characteristics make LLM-based forecasting especially appealing for integration into mobile health platforms, wearable CGM devices, and real-time insulin delivery systems.

6.5 Limitations of the Current Study

One key limitation of this project is the risk of data contamination: publicly released LLMs may have encountered fragments of benchmark CGM datasets during pre-training, which could artificially inflate their zero-shot performance. While CGM records are relatively rare in large-scale pre-training corpora—especially compared to domains like finance or meteorology—the possibility cannot be fully ruled out. To address this concern, future work should consider evaluating models on private or held-out clinical datasets to ensure an unbiased assessment.

A second limitation arises from the inherent non-determinism of LLM output. Due to stochastic sampling, repeated forecasts can vary substantially. Although averaging predictions across multiple runs improves stability for benchmarking, this approach may prove cumbersome or inconsistent in real-time clinical applications. One straightforward way to reduce run-to-run variability is to constrain the decoding process itself. A practical solution involves constraining the decoding process—such as lowering the temperature, applying strict top-k or top-p filtering, or adopting deterministic strategies like greedy or beam search. These methods can yield more consistent forecasts, though they may reduce output diversity and impair performance.

6.6 Future Research Directions

Looking ahead, several directions may further enhance the effectiveness of LLM-based CGM forecasting. One avenue involves systematic tuning of generation hyperparameters—such as temperature, top-k, top-p, and logit bias—to better align the model's output distribution

with the dynamics of CGM data. Careful adjustment of these parameters could significantly enhance predictive accuracy by reducing noise, improving stability, and tailoring the model's behavior to glycemic trends.

Another promising direction would be to explore privacy-preserving personalization strategies through federated fine-tuning(Ye et al., 2024). This approach would enable each patient's device to locally adapt a distilled LLM using their own CGM data—including insulin dosing history, meal logs, and daily routines—without transmitting sensitive information to a centralized server. Instead, only model updates would be shared, allowing a global model to improve over time while respecting patient privacy. Such personalized adaptation could significantly boost performance on rare or outlier patterns that are difficult to capture with population-level training alone.

CHAPTER 7

Conclusion

In this thesis, we have explored the zero-shot forecasting capabilities of large language models (LLMs) for continuous glucose monitoring (CGM) data, employing two distinct paradigms: a Direct-Prompt framework and a Retrieval-Augmented Generation (RAG) framework. Across five publicly available CGM datasets, we systematically evaluated each approach's ability to incorporate patient-specific contextual information and deliver accurate forecasts, benchmarking performance against classical time series models (ARIMA, linear regression), a hybrid model (LatentODE), and modern deep learning approaches (Transformer).

Our experiments yield several notable insights. First, even without retrieval augmentation, LLMs, especially smaller distilled models like GPT-4o-mini, demonstrate strong zero-shot forecasting performance. In settings with small sample sizes and high time series variability—such as the Broll2021 dataset—LLMs significantly outperform traditional baselines, highlighting their ability to leverage pre-trained general knowledge for robust forecasting. Moreover, in moderately to larger-sized datasets like Hall2018 and Weinstock2016, direct-prompt LLMs perform on par with, and occasionally surpass, complex deep learning models that required extensive training and tuning.

Second, augmenting LLMs with relevant historical CGM sequences via RAG leads to further improvements in accuracy for larger models like GPT-4o. Retrieval metrics such as mean absolute percentage error (MAPE) and pearson correlation consistently yielded the strongest gains, while euclidean distance performed best in shorter input-length scenarios.

Notably, we also discovered situations in which LLM-based methods underperform. For instance, on datasets such as Colas2019 and Dubosson2018, linear regression methods continue to outperform more complex models, including LLM-based approaches. Models with complex architectures—including LLMs—can still suffer from overfitting and noise-induced prediction instability.

Taken together, our findings position LLMs as a promising and versatile tool for personalized CGM data forecasting. Unlike traditional time series models, LLMs can seamlessly

integrate contextual cues (e.g., diabetes type, meal intake, insulin dosing) into their predictions through natural language prompts. Moreover, zero-shot deployment eliminates the computational burden of training, feature engineering, and hyperparameter tuning—making this approach practical and scalable for real-world clinical or consumer use.

In conclusion, LLMs—augmented through retrieval or guided by contextual prompting—do not replace classical forecasting models, but instead offer a flexible and extensible complement. As LLM capabilities continue to evolve, their potential to transform CGM analytics becomes increasingly within reach.

CHAPTER 8

Ethical Considerations and Data Privacy

The development of predictive models using Continuous Glucose Monitoring (CGM) data carries significant ethical responsibility. In this work, we remain firmly committed to protecting patient privacy and promoting responsible data stewardship at every stage of the research process.

All raw CGM datasets used in this study were obtained from publicly available sources and were originally collected under robust informed consent protocols. Participants were fully briefed on the nature of data collection, the scope of its intended use in research, and their right to withdraw participation at any time. Furthermore, each dataset underwent rigorous de-identification procedures. Personally identifiable information—such as names, contact details, or geographic locations—was removed or replaced with randomized subject identifiers to minimize the risk of re-identification.

Looking ahead, we recognize the importance of ongoing collaboration with patients, clinicians, and other stakeholders as LLM-based forecasting tools are refined and deployed. We advocate for the continued development of privacy-preserving methodologies—such as federated learning—that can enable model personalization without exposing raw patient data.

APPENDIX A

Full Retrieval-Augmented Generation Results

Color key:
Blue entries outperform the respective GPT-4o/4o mini direct-prompt baseline. Red entries underperform it.

Table A.1: Prediction accuracy for all RAG configurations

Metric	L	Model	Dataset	RMSE	MAE
correlation	72	gpt-4o-mini	Colas	9.0108	7.9444
correlation	144	gpt-4o-mini	Colas	7.4734	6.8333
correlation	72	gpt-4o	Colas	7.1918	6.2222
correlation	144	gpt-4o	Colas	8.1029	7.0556
correlation	72	gpt-4o-mini	Dubosson	27.2488	25.7806
correlation	72	gpt-4o	Dubosson	15.8910	14.1778
correlation	144	gpt-4o	Dubosson	22.3182	18.9167
correlation	144	gpt-4o-mini	Dubosson	13.4035	11.1972
correlation	72	gpt-4o-mini	Hall	7.2085	6.1528
correlation	144	gpt-4o-mini	Hall	6.3740	5.5833
correlation	72	gpt-4o	Hall	6.2078	5.1667
correlation	144	gpt-4o	Hall	6.8527	6.0278
correlation	72	gpt-4o	Broll	10.1909	8.7500
correlation	72	gpt-4o-mini	Broll	16.9103	16.0556
correlation	144	gpt-4o	Broll	4.7113	3.9722
correlation	144	gpt-4o-mini	Broll	7.2685	6.4444
correlation	72	gpt-4o-mini	Weinstock	17.2302	14.3681
correlation	144	gpt-4o-mini	Weinstock	17.4257	15.2222

Continued on next page

Table A.1 (continued)

Metric	L	Model	Dataset	RMSE	MAE
correlation	72	gpt-4o	Weinstock	13.7198	12.1597
correlation	144	gpt-4o	Weinstock	16.2534	13.3333
euclidean	72	gpt-4o-mini	Colas	8.5521	7.7500
euclidean	144	gpt-4o-mini	Colas	8.1383	7.1389
euclidean	72	gpt-4o	Colas	7.7166	6.8889
euclidean	72	gpt-4o	Dubosson	18.5915	15.6611
euclidean	72	gpt-4o-mini	Dubosson	15.5520	12.5194
euclidean	144	gpt-4o-mini	Dubosson	17.3875	14.5639
euclidean	72	gpt-4o-mini	Hall	8.4423	7.4583
euclidean	144	gpt-4o-mini	Hall	6.8689	5.7361
euclidean	72	gpt-4o	Hall	5.4795	4.5417
euclidean	72	gpt-4o	Broll	8.9038	7.3194
euclidean	72	gpt-4o-mini	Broll	17.5454	16.4444
euclidean	144	gpt-4o-mini	Broll	14.4627	13.9722
euclidean	72	gpt-4o-mini	Weinstock	17.9959	15.9722
euclidean	144	gpt-4o-mini	Weinstock	16.8292	14.9444
euclidean	72	gpt-4o	Weinstock	13.9505	12.0903
MAPE	72	gpt-4o-mini	Colas	8.8652	7.9444
MAPE	144	gpt-4o-mini	Colas	8.2971	7.1667
MAPE	72	gpt-4o	Colas	7.1828	5.9167
MAPE	144	gpt-4o	Colas	7.7919	6.6389
MAPE	72	gpt-4o-mini	Dubosson	19.0894	15.6639
MAPE	144	gpt-4o-mini	Dubosson	17.4042	15.8611
MAPE	72	gpt-4o	Dubosson	15.7254	14.8306
MAPE	144	gpt-4o	Dubosson	12.2333	9.4000
MAPE	72	gpt-4o-mini	Hall	7.5734	6.4361
MAPE	144	gpt-4o-mini	Hall	6.8676	5.7556
MAPE	72	gpt-4o	Hall	5.8542	5.0694
MAPE	144	gpt-4o	Hall	7.4927	6.5000
MAPE	72	gpt-4o	Broll	16.3204	13.6389
MAPE	72	gpt-4o-mini	Broll	12.1735	10.4722
MAPE	144	gpt-4o	Broll	4.9251	4.0347

Continued on next page

Table A.1 (continued)

Metric	L	Model	Dataset	RMSE	MAE
MAPE	144	gpt-4o-mini	Broll	7.8381	7.4167
MAPE	72	gpt-4o-mini	Weinstock	16.7541	15.1181
MAPE	144	gpt-4o	Weinstock	17.9150	14.6417
MAPE	144	gpt-4o-mini	Weinstock	15.4668	13.6250
MAPE	72	gpt-4o	Weinstock	15.5106	12.8194

BIBLIOGRAPHY

Ansari, A. F., L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. P. Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, H. Wang, M. W. Mahoney, K. Torkkola, A. G. Wilson, M. Bohlke-Schneider, and Y. Wang (2024, November). Chronos: Learning the Language of Time Series.

Box, G. E. P. and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control* (Revised Edition ed.). San Francisco: Holden Day.

Broll, S., J. Urbanek, D. Buchanan, E. Chun, J. Muschelli, N. M. Punjabi, and I. Gaynanova (2021). Interpreting blood GLUcose data with R package iglu. *16*(4), e0248560.

Carlson, A. L., D. M. Mullen, and R. M. Bergenstal (2017, May). Clinical Use of Continuous Glucose Monitoring in Adults with Type 2 Diabetes. *Diabetes Technology & Therapeutics 19*(S2), S–4–S–11.

Chrzanowski, J., S. Grabia, A. Michalak, A. Wielgus, J. Wykrota, B. Mianowska, A. Szadkowska, and W. Fendler (2022, November). GlyCulator 3.0: A Fast, Easy-to-Use Analytical Tool for CGM Data Analysis, Aggregation, Center Benchmarking, and Data Sharing. *Diabetes Care 46*(1), e3–e5.

Colás, A., L. Vigil, B. Vargas, D. Cuesta-Frau, and M. Varela (2019). Detrended Fluctuation Analysis in the prediction of type 2 diabetes mellitus in patients at risk: Model optimization and comparison with other metrics. *PloS One 14*(12), e0225817.

Dong, Q., L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, B. Chang, X. Sun, L. Li, and Z. Sui (2024, October). A Survey on In-context Learning.

Dubosson, F., J.-E. Ranvier, S. Bromuri, J.-P. Calbimonte, J. Ruiz, and M. Schumacher (2018, January). The open D1NAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management. *Informatics in Medicine Unlocked 13*, 92–100.

Gao, Y., Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang (2024, March). Retrieval-Augmented Generation for Large Language Models: A Survey.

Gecili, E., R. Huang, J. C. Khoury, E. King, M. Altaye, K. Bowers, and R. D. Szczesniak (2020). Functional data analysis and prediction tools for continuous glucose-monitoring studies. *Journal of Clinical and Translational Science 5*(1), e51.

Gruver, N., M. Finzi, S. Qiu, and A. G. Wilson (2024, August). Large Language Models Are Zero-Shot Time Series Forecasters.

Hall, H., D. Perelman, A. Breschi, P. Limcaoco, R. Kellogg, T. McLaughlin, and M. Snyder (2018, July). Glucotypes reveal new patterns of glucose dysregulation. *PLOS Biology 16*(7), e2005143.

Jin, M., S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, and Q. Wen (2023, October). Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela (2021, April). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.

Oser, S. M. and T. K. Oser (2024, January). Medicare Coverage of Continuous Glucose Monitoring – 2023 Updates. *Family Practice Management 31*(1), 17–18.

Prendin, F., S. Del Favero, M. Vettoretti, G. Sparacino, and A. Facchinetti (2021, January). Forecasting of Glucose Levels and Hypoglycemic Events: Head-to-Head Comparison of Linear and Nonlinear Data-Driven Algorithms Based on Continuous Glucose Monitoring Data Only. *Sensors 21*(5), 1647.

Radford, A. and K. Narasimhan (2018). Improving Language Understanding by Generative Pre-Training.

Rodbard, D. (2016, February). Continuous Glucose Monitoring: A Review of Successes, Challenges, and Opportunities. *Diabetes Technology & Therapeutics 18*(S2), S2–3–S2–13.

Rubanova, Y., R. T. Q. Chen, and D. Duvenaud (2019, July). Latent ODEs for Irregularly-Sampled Time Series.

Sergazinov, R., E. Chun, V. Rogovchenko, N. Fernandes, N. Kasman, and I. Gaynanova (2024). GlucoBench: Curated List of Continuous Glucose Monitoring Datasets with Prediction Benchmarks.

Song, Y., G. Wang, S. Li, and B. Y. Lin (2024, July). The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism.

Tire, K., E. O. Taga, M. E. Ildiz, and S. Oymak (2024, November). Retrieval Augmented Time Series Forecasting.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample (2023, February). LLaMA: Open and Efficient Foundation Language Models.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2023, August). Attention Is All You Need.

Vigers, T., C. L. Chan, J. Snell-Bergeon, P. Bjornstad, P. S. Zeitler, G. Forlenza, and L. Pyle (2019). Cgmanalysis: An R package for descriptive analysis of continuous glucose monitor data. *PloS One 14*(10), e0216851.

Weinstock, R. S., S. N. DuBose, R. M. Bergenstal, N. S. Chaytor, C. Peterson, B. A. Olson, M. N. Munshi, A. J. Perrin, K. M. Miller, R. W. Beck, D. R. Liljenquist, G. Aleppo, J. B. Buse, D. Kruger, A. Bhargava, R. S. Goland, R. C. Edelen, R. E. Pratley, A. L. Peters, H. Rodriguez, A. J. Ahmann, J.-P. Lock, S. K. Garg, M. R. Rickels, I. B. Hirsch, and for the T1D Exchange Severe Hypoglycemia in Older Adults With Type 1 Diabetes Study Group (2016, December). Risk Factors Associated With Severe Hypoglycemia in Older Adults With Type 1 Diabetes. *Diabetes Care 39*(4), 603–610.

Williams, A. R., A. Ashok, É. Marcotte, V. Zantedeschi, J. Subramanian, R. Riachi, J. Requeima, A. Lacoste, I. Rish, N. Chapados, and A. Drouin (2025, February). Context is Key: A Benchmark for Forecasting with Essential Textual Information.

Ye, R., W. Wang, J. Chai, D. Li, Z. Li, Y. Xu, Y. Du, Y. Wang, and S. Chen (2024, February). OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen (2025, March). A Survey of Large Language Models.