Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master Thesis in Bioengineering

Automatic Analysis of Continuous Glucose Monitoring Data in Type 1 Diabetes Using Large Language Models in Clinical Practice

Master Candidate

Saiok Miah

Student ID 2071912

Supervisor

Prof. Giacomo Cappon

University of Padova

Co-supervisor

Dott. Luca Cossu

University of Padova

Academic Year
2023/2024

To the ones
who always believed in me

Abstract

Continuous glucose monitoring (CGM) has revolutionized the management of Type 1 Diabetes (T1D) by providing real-time data on glucose levels. However, the complexity of CGM data presents challenges for clinical interpretation and actionable insights. This thesis explores the application of large language models (LLMs) to facilitate the analysis of CGM data, aiming to enhance clinical decision-making in T1D management. The study seeks to summarize the Ambulatory Glucose Profile (AGP), which contains essential metrics and patterns from CGM readings, into ready-to-use summaries. The BioMistral model from Hugging Face, a cutting-edge LLM designed for biomedical applications, is employed for this purpose. By extracting relevant clinical metrics from glucose data and designing ad-hoc prompts for the LLM, we aim to replicate the clinical analysis of the AGP and present it in a simplified manner. With the help of clinical partners, we evaluated the effectiveness and safety of the summaries for clinical practice. This automated analysis system can improve the accuracy and efficiency of glucose management, offering clinicians a powerful tool to support patient care. The integration of BioMistral into clinical practice has the potential to enhance T1D management, providing a scalable, data-driven solution for better interpretation of patient outcomes.

Sommario

Il monitoraggio continuo del glucosio (CGM) ha rivoluzionato la gestione del diabete di tipo 1 (T1D) fornendo dati in tempo reale sui livelli di glucosio. Tuttavia, la complessità dei dati CGM pone sfide per l'interpretazione clinica e le informazioni attuabili. Questa tesi esplora l'applicazione di modelli linguistici di grandi dimensioni (LLMs) per facilitare l'analisi dei dati CGM, con l'obiettivo di migliorare il processo decisionale clinico nella gestione del T1D. Lo studio cerca di riassumere il profilo ambulatoriale del glucosio (AGP), che contiene metriche e modelli essenziali dalle letture di CGM, in riassunti pronti all'uso. Il modello BioMistral di Hugging Face, un LLM all'avanguardia progettato per applicazioni biomediche, viene impiegato a tal fine. Estraendo le metriche cliniche rilevanti dai dati del glucosio e progettando i prompt ad hoc per il LLM, miriamo a replicare l'analisi clinica del AGP e a presentarla in un modo semplificato. Con l'aiuto di partner clinici, abbiamo valutato l'efficacia e la sicurezza dei sommari per la pratica clinica. Questo sistema di analisi automatizzato può migliorare l'accuratezza e l'efficienza della gestione del glucosio, offrendo ai medici uno strumento potente per supportare la cura dei pazienti. L'integrazione di BioMistral nella pratica clinica ha il potenziale per migliorare la gestione del T1D, fornendo una soluzione scalabile e basata sui dati per una migliore interpretazione dei risultati dei pazienti.

Contents

List of Figures xi

List of Tables xiii

List of Acronyms xix

1 Introduction **1**

1.1 Diabetes 1

1.1.1 Glucose Monitoring 2

1.2 Data management and interpretation in T1D 5

1.2.1 Ambulatory Glucose Profile 6

1.2.2 Components of the Ambulatory Glucose Profile 6

1.3 Clinical Utility of AGP 9

1.4 Challenges in data interpretation and analysis 9

1.5 Structure and aim of the thesis 10

2 State of the Art **13**

2.1 Background 13

2.2 Overview of Developments with New Technological Advancements 14

2.3 How GPT Models Work 15

2.3.1 Training and Fine-Tuning 16

2.3.2 Model Architecture 16

2.3.3 Prompt Design and Prompt Engineering 18

3 Framework and Data Processing **21**

3.1 Data Processing Pipeline 21

3.2 Tidepool Dataset and Processing with AGATA 23

3.2.1 Dataset: Tidepool CGM Data 23

3.2.2 Preprocessing with the AGATA Library 23
3.3 BioMistral: A Medical Large Language Model 24
3.4 Testing Process with GPT Models on AGP Summaries 25

4 Summarizing the AGP data and metrics using BioMistral 27
4.1 The Role of GPT in Medical Explanations 28
4.1.1 GPT Integration 28
4.1.2 Advantages of BioMistral 28
4.2 Prompt Engineering for GPT - BioMistral 28
4.2.1 Iterative Improvement of Prompts Based on Test Results . 29
4.3 Analyzing the performance and effectiveness of AGP summaries 34
4.4 Evaluation Methodology of Responses 35

5 Results and Discussion 37
5.1 Results overview 37
5.1.1 AGP Response Analysis 37
5.2 Evaluation Process 41
5.3 Discussion of Results 43
5.3.1 Enhancing Efficiency in Clinical Workflow 44
5.3.2 Significance of Time in Range and Glycemic Pattern Detection 44
5.3.3 Comparative Analysis of Model Limitations 44
5.3.4 Importance of Clinical Oversight and Contextual Understanding 45
5.3.5 The Role of Prompt Engineering in Optimizing Performance 45
5.3.6 Advancing Model Sensitivity Through Contextual Cues . 45

6 Conclusions and Future Work 47
6.1 Conclusion 47
6.1.1 Key Findings 47
6.1.2 Limitations 48
6.1.3 Concluding Remarks 49
6.2 Future Directions and Clinical Integration 49

References 51

Acknowledgments 57

List of Figures

1.1 Example of self monitoring of blood glucose 3
1.2 Example of continuous glucose monitoring 4
1.3 Ambulatory Glucose Profile 7

2.1 Transformer architecture for GPT models 17

3.1 Data Processing Pipeline 22

4.1 Linkert scale for evaluating responses 36

5.1 AGP of patient 1 38
5.2 AGP of patient 7 39
5.3 AGP of patient 8 40
5.4 Graph of the results 43

List of Tables

5.1 Scores for BioMistral outputs 42
5.2 Scores for ChatGPT outputs 42

List of Acronyms

AGATA Automated Glucose dATa Analysis

AGP Ambulatory Glucose Profile

AI Artificial Intelligence

CV coefficient of variation

CGM Continuous Glucose Monitoring

cDSS Clinical Decision Support Systems

EHR Electronic Health Record

GDPR General Data Protection Regulation

GMI Glucose Management Indicator

GPT Generative Pre-trained Transformer

LLM Large Language Model

LLMs Large Language Models

ML Machine Learning

MDI Multiple Daily Injections

NLP Natural Language Processing

RL Reinforcement Learning

SMBG Self-Monitoring of Blood Glucose

T1D Type 1 Diabetes

T2D Type 2 Diabetes

TAR Time Above Range

TBR Time Below Range

TIR Time in Range

1 Introduction

1.1 Diabetes

Diabetes is a chronic health condition characterized by abnormally high or insufficient levels of blood sugar (glucose). This disorder occurs when the body is either unable to produce enough insulin or cannot utilize the insulin it does produce effectively. Insulin is a crucial hormone, produced by the pancreas, that helps regulate blood glucose levels by facilitating the transport of glucose into cells to be converted into energy [1]. Inadequate insulin function results in glucose accumulating in the bloodstream, leading to various health issues. Diabetes is primarily classified into three major types: Type 1 Diabetes (T1D), Type 2 Diabetes (T2D), and gestational diabetes [2].
T1D is an autoimmune condition typically diagnosed in children and young adults, where the immune system mistakenly attacks and destroys insulin producing cells in the pancreas [3]. T2D is more common and is caused by genetic and lifestyle factors, leading to insulin resistance. It usually develops in adults, but more young people are getting it [4]. Gestational diabetes is a temporary condition that occurs during pregnancy and elevates the risk of developing T2D later in life.

Diagnosis of diabetes involves a range of tests such as fasting blood glucose levels, oral glucose tolerance tests, and hemoglobin A1c measurements, which help monitor long-term glucose control [1, 5]. Effective management of diabetes is crucial to avoid both acute and chronic complications associated with imbalances in blood glucose levels. Such complications include hyperglycemia, or elevated blood glucose, can lead to long-term complications such as organ damage, particularly affecting the eyes, kidneys, and cardiovascular system [6]. On the contrary, hypoglycemia, or reduced blood glucose, can cause immediate life-threatening situations such as seizures, loss of consciousness, and even death if not promptly addressed [7]. Therefore, both patients and healthcare providers must manage these fluctuations carefully to ensure optimal glycemic control.

1.1.1 Glucose Monitoring

Monitoring blood glucose levels accurately and regularly is essential for managing diabetes effectively. Several technologies are available that help both patients and healthcare providers maintain good glycemic control. These technologies provide valuable information for adjusting insulin doses, diet, and physical activity, improving overall disease management [8].

Self-Monitoring of Blood Glucose

Self-Monitoring of Blood Glucose (SMBG) is one of the most widely used technologies for managing diabetes. This conventional technique involves using a blood glucose meter to measure glucose levels in capillary blood, typically obtained from a finger prick. SMBG provides an instantaneous glucose reading, allowing patients to make real-time adjustments to their insulin doses, dietary intake, or physical activity. While it is useful for spot checks and post-prandial glucose measurements, SMBG is limited by its discrete nature, as it only provides snapshots of glucose levels at specific times and may miss fluctuations between tests [9]. Despite its limitations, SMBG remains state of the art and reference tool, particularly for individuals managing their diabetes with multiple daily insulin injections.

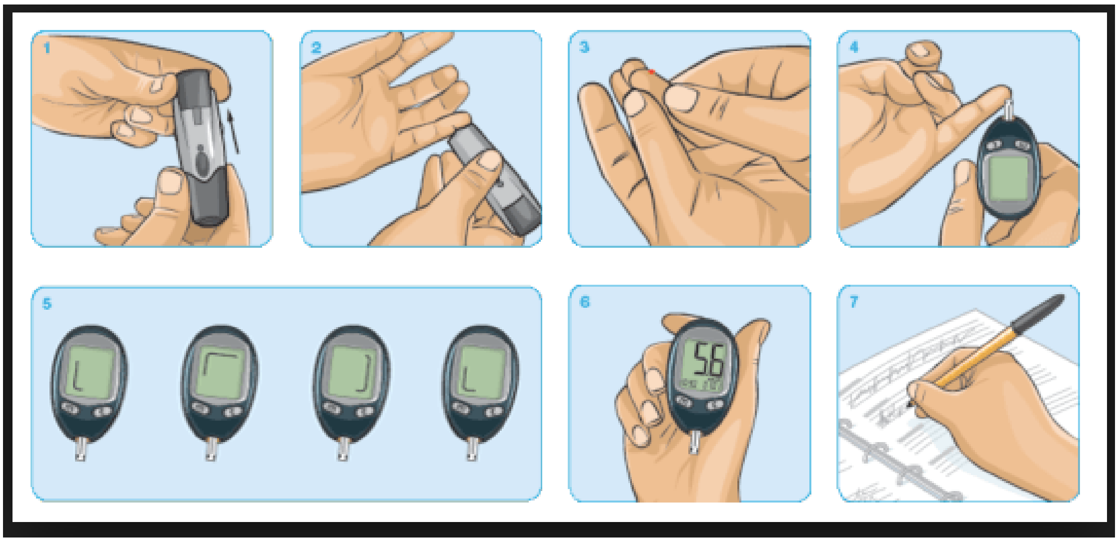

Figure 1.1: Self monitoring of blood glucose, steps to measure blood glucose: 1. Prepare the puncture device. 2. Prick a finger with the device. 3. Prick your finger to get a drop of blood. 4. Apply the blood to the strip that is first inserted into the meter. 5. Wait for the counter to display the result. 6. Check glucose reading. 7. Record the result in a log.

Continuous Glucose Monitoring

Continuous Glucose Monitoring (CGM) represents a significant advancement in diabetes care. CGM systems employ a small sensor placed under the skin to measure glucose levels in the interstitial fluid continuously. These devices provide real-time glucose readings, allowing patients and clinicians to monitor glucose trends over time [10]. CGM systems typically include alerts for hyperglycemia or hypoglycemia, enabling more proactive management of glucose levels. This continuous stream of data is especially beneficial for individuals with Type 1 diabetes, those prone to frequent hypoglycemia, or patients striving for tight glucose control [11]. CGM significantly reduces the frequency and severity of both hyper- and hypoglycemic events, offering patients and healthcare providers a more comprehensive understanding of daily glucose variations.

Figure 1.2: Continuous glucose monitoring (CGM): Real-time blood glucose data displayed on a smartphone synced with an arm sensor.

Unlike SMBG, which only provides single readings, CGM gives a detailed picture of blood sugar levels throughout the day. It can also alert users when their blood sugar is too high or too low, which helps them take action before serious problems arise [12]. This continuous data helps patients and doctors make better decisions about insulin use and lifestyle changes, leading to fewer complications and better control of diabetes.

1.2 Data management and interpretation in T1D

The use of these glucose monitoring technologies has transformed diabetes management by empowering patients to take a more active role in their care [13, 14]. In particular, CGM provide patients with the ability to closely monitor and respond to their glucose levels throughout the day, enhancing overall glycemic control and reducing the risk of acute complications such as diabetic ketoacidosis or severe hypoglycemia [15]. The integration of CGM data into clinical workflows also supports healthcare providers in developing more personalized and effective treatment plans for their patients [16].

Managing diabetes effectively requires not only understanding the underlying mechanisms of the disease but also maintaining tight control over blood glucose levels to avoid complications such as hyperglycemia and hypoglycemia [15]. While traditional monitoring methods like SMBG have been instrumental in helping patients make real-time adjustments, they often fail to provide a complete picture of daily glucose fluctuations.

In contrast, CGM systems offer a more comprehensive view of glucose trends by continuously tracking blood sugar levels. However, interpreting the vast amount of data generated by CGM systems can be challenging for both patients and healthcare providers. This is where the Ambulatory Glucose Profile (AGP) plays a crucial role.
The AGP consolidates complex CGM data into a clear, standardized format, making it easier for clinicians and patients to identify patterns and make informed decisions about treatment. By focusing on key metrics such as Time in Range (TIR) and glucose variability, AGP provides actionable insights that can significantly improve glycemic control and reduce the risk of complications associated with diabetes.

1.2.1 Ambulatory Glucose Profile

The AGP is a standardized and comprehensive graphical representation of glucose data collected via CGM systems. It is increasingly recognized as a valuable tool in the management of diabetes, providing both patients and healthcare professionals with a clear, concise, and actionable summary of glucose patterns over a specified period, typically 14 days [16]. Structured around a daily 24-hour glucose profile, the AGP aggregates and displays data in a simplified overlay format, facilitating the identification of recurring glucose trends and variations. By consolidating complex glucose data into an easily interpretable format, the AGP highlights critical metrics such as TIR, glucose variability, and periods of hyperglycemia and hypoglycemia. This chapter explores the utility of AGP in diabetes care, emphasizing its role in optimizing glycemic control and minimizing the risks of complications associated with diabetes [5, 13].

1.2.2 Components of the Ambulatory Glucose Profile

The AGP report consists of several key elements that collectively offer a detailed overview of the patient's glycemic status [13, 16]:

Median Glucose Levels and Percentiles

The AGP plot graphically represents the median glucose levels alongside the 25th and 75th percentiles (interquartile range) and the 5th and 95th percentiles. This visual distribution of glucose data over a typical 24-hour period helps identify the degree of glucose variability and the consistency of glucose control.

Time in Range

TIR is one of the most critical metrics in the AGP report, indicating the percentage of time a patient's glucose levels remain within the target range (70-180 mg/dL). A higher TIR is associated with a reduced risk of diabetes-related complications, particularly microvascular complications [5].
These metrics provide insights into the duration of hyperglycemia Time Above Range (TAR) and hypoglycemia Time Below Range (TBR), offering a clear picture of glucose excursions outside the target range. Effective diabetes management aims to minimize TAR and TBR to avoid long-term complications and acute hypoglycemic episodes.

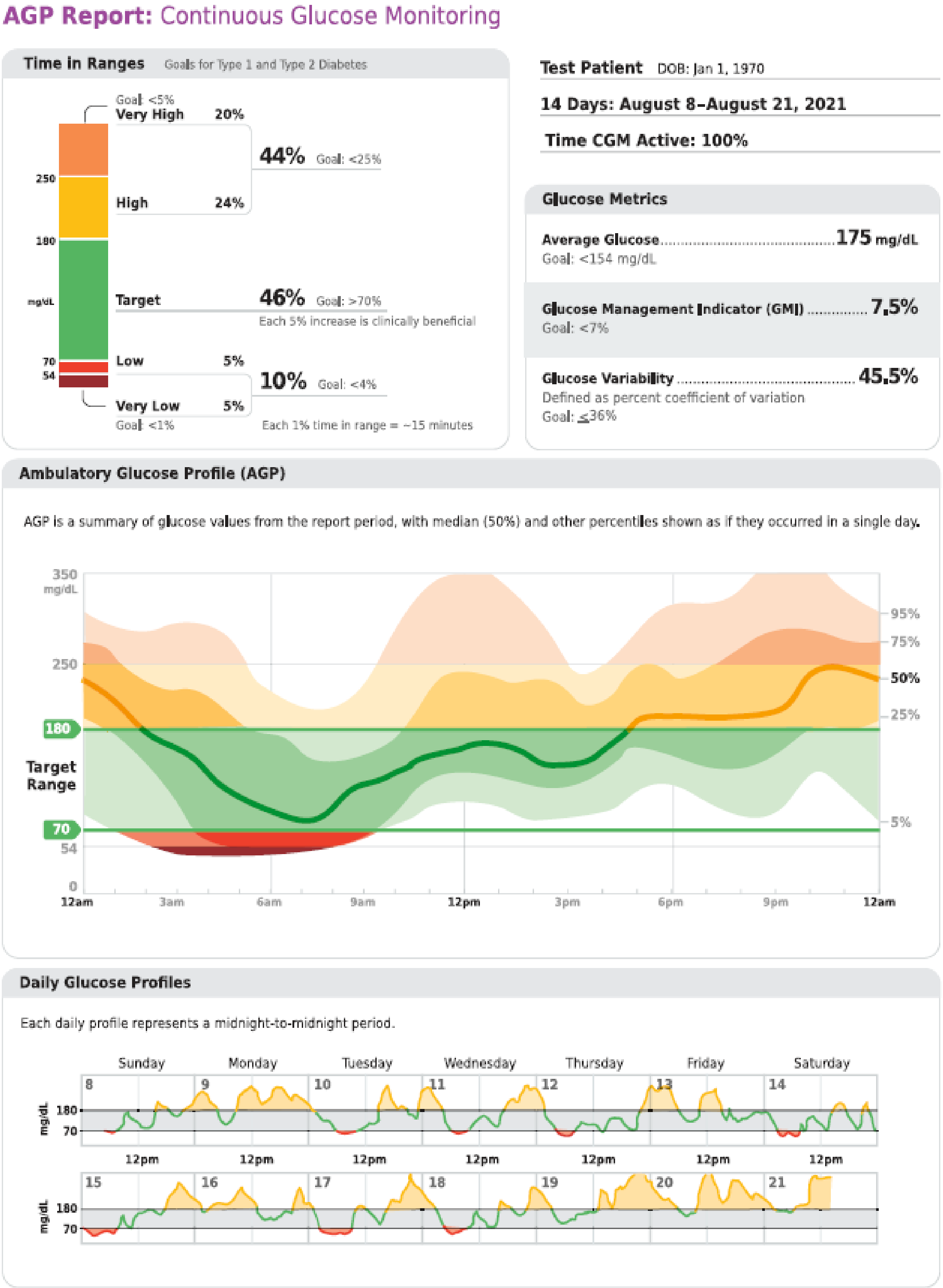

Figure 1.3: AGP summarizing two weeks of CGM data. The solid green line represents the median glucose, shaded areas show variability (25th–75th and 5th–95th percentiles), and the time in range is highlighted. The daily trends below help identify patterns like hypoglycemia or hyperglycemia.

Glucose Variability

The AGP provides measures of glucose variability, a fundamental parameter in diabetes management, as higher variability is often associated with an increased risk of hypoglycemia and other glucose-related complications. Glucose variability is typically expressed through the coefficient of variation coefficient of variation (CV), a critical indicator for optimizing diabetes management due to its ability to identify extreme glucose fluctuations and assess the stability of glycemic control [17]. The coefficient of variation is calculated as follows:

$$CV = \frac{\sigma}{\mu} \times 100 \tag{1.1}$$

where σ is the standard deviation of glucose levels and μ is the mean glucose level [mg/dL]. In the literature, a CV of less than 36% is often considered acceptable to reduce the risk of hypoglycemia and ensure a safer, more stable glycemic management [18].

Glucose Management Indicator (GMI)

The Glucose Management Indicator (GMI) is an estimate of HbA1c based on the mean glucose levels recorded by Continuous Glucose Monitoring (CGM) over the monitoring period, providing a quick and practical assessment of long-term glycemic control [19]. HbA1c, or glycated hemoglobin, is a measure of the average blood glucose level over the past two to three months, reflecting the percentage of glucose bound to hemoglobin in red blood cells. It is a key indicator for assessing long-term glucose control in diabetes management.
The GMI is calculated using the following formula:

$$GMI = 3.31 + 0.02392 \times \mu \tag{1.2}$$

The coefficient 0.02392 translates the mean glucose level into an equivalent HbA1c percentage, while the constant 3.31 adjusts the calculation to align with laboratory-based HbA1c readings, ensuring the resulting value is a reliable approximation of HbA1c. This metric allows healthcare providers to estimate HbA1c without the delay of laboratory testing, facilitating timely and targeted therapeutic decisions. As such, the GMI has become an important tool for clinicians and patients, offering an immediate snapshot of glycemic control that supports more proactive diabetes management [19].

1.3 Clinical Utility of AGP

The AGP is not only a data summary tool but also a valuable clinical decision-making aid. It enables the identification of glucose patterns that may not be evident from HbA1c values alone, such as nocturnal hypoglycemia and post-prandial hyperglycemia [12]. The AGP provides a visual representation of glucose levels throughout the day, which facilitates a more nuanced understanding of glycemic control, enabling healthcare providers to make more personalized adjustments to therapy [20].

The AGP plays a crucial role in achieving individualized glycemic goals. For instance, a target of greater than 70% TIR is often recommended for most non-pregnant adults, with less stringent targets for older adults or those with significant comorbidities [12]. Through its visual and intuitive interface, the AGP assists clinicians and patients in collaborating to set realistic goals and monitor progress, thereby supporting a balanced approach to glycemic control while minimizing the risk of hypoglycemia [21].

1.4 Challenges in data interpretation and analysis

While the AGP offers numerous advantages in presenting CGM data, its use in clinical settings also presents several challenges. Effective diabetes management requires both patients and healthcare providers to interpret complex CGM data accurately, making informed decisions regarding insulin, diet, and lifestyle adjustments [21]. However, the potential of CGM data remains underutilized in clinical practice, largely due to challenges in data interpretation, time constraints, and the lack of accessible tools that simplify this information for practical application [13].

Moreover, the AGP, while valuable in summarizing glucose trends, has certain limitations. For instance, it does not capture specific glucose patterns, such as after meals or low blood sugar episodes during the night, which are essential for optimizing insulin treatment [22]. Additionally, the AGP does not account for contextual factors like mealtime and physical activity, both of which significantly influence blood glucose levels and treatment outcomes [20]. Consequently, the AGP, though helpful, may overlook critical details necessary for truly individualized care.

Most Clinical Decision Support Systems (cDSS) for T1D management include

two primary components: data analysis tools and treatment recommendation systems. The AGP is still one of the most commonly used tools, but even this system cannot fully handle the large amount of information from CGM devices [12]. Currently, many cDSS systems provide generalized recommendations derived from standard clinical guidelines. However, these suggestions may not be personalized and may not address the specific needs of every patient [23]. Furthermore, the majority of existing cDSS do not provide detailed explanations of their recommendations, reducing their utility and acceptance in daily clinical practice [24].
Managing T1D is a complex process that requires consideration of various factors, including insulin dosage, timing, and patient adherence. Unfortunately, most cDSS tools do not sufficiently address this complexity, forcing clinicians to rely on their experience and judgment in making treatment decisions [25]. It is evident that more advanced, automated cDSS systems are needed to analyze and present CGM data in an accessible and actionable format. Such systems would enable clinicians to make quicker and more informed decisions, which would ultimately improve diabetes management and patient outcomes [13].

1.5 Structure and aim of the thesis

The objective of this thesis is to develop a component of an automated cDSS that utilizes a Large Language Model (LLM), specifically BioMistral, to summarize AGP data in a concise and accurate manner. The system is designed to provide clinicians with key insights, highlighting critical areas requiring further evaluation, while alleviating the burden of manually analyzing complex AGP reports. By automating this repetitive and time-intensive process, the cDSS aims to enable clinicians to dedicate more time to high-value activities, such as personalizing treatment plans for individual patients, ultimately enhancing the quality of care in the management of type 1 diabetes.

This thesis begins by reviewing the current state of cDSS in diabetes management, examining their limitations and exploring how advanced AI-driven models, including recent innovations such as Generative Pre-trained Transformer (GPT), can overcome these challenges by interpreting CGM data and supporting personalized treatment, with Chapter 2 focusing specifically on the state of the art in conventional cDSS platforms and the application of machine learning

to T1D management.
In Chapter 3, the framework and data processing workflow are presented, with detailed steps for analyzing CGM data, calculating key metrics, and generating AGP plots explained. Advanced language models like BioMistral and ChatGPT are integrated to create concise clinical summaries, which ensure the system is both practical and ethically sound for supporting diabetes management.
The application of BioMistral in summarizing AGP data for pediatric patients with Type 1 diabetes is explored in Chapter 4, demonstrating its ability to translate complicated glucose metrics into actionable insights. The chapter focuses on the iterative development of prompts, highlighting techniques like roleplay and structured instructions to refine the model's clinical relevance. We analyze how BioMistral transforms CGM data into tailored therapy recommendations, comparing its outputs with those of ChatGPT in terms of informativeness, accuracy, and utility for healthcare providers.
The performance of the system in producing clinically relevant insights is evaluated by comparing BioMistral's outputs with those of ChatGPT models in chapter 5. Each model's accuracy, safety, completeness, and clinical utility are analyzed in this chapter through detailed expert feedback from diabetologists. The findings show how the models differ in their ability to generate summaries that are safe, accurate, and informative, which provides a foundation for improvement and confirms the feasibility of using advanced LLMs in clinical settings.
Chapter 6 concludes the thesis by summarizing the main findings, emphasizing the impact of LLM technology on diabetes management, and discussing the broader implications of integrating such systems in clinical workflows. The chapter also identifies future developments that could enhance the system's performance.

2 State of the Art

This chapter provides a comprehensive review of the current state of the art in decision support systems developed for Type 1 diabetes management. We focus on both traditional Clinical Decision Support Systems (cDSS) systems and the latest advancements incorporating Artificial Intelligence (AI) and Machine Learning (ML) techniques. Emphasis is placed on how these emerging technologies enhance the ability of cDSS to interpret CGM data, providing clinicians with actionable insights for personalized patient care. Furthermore, we explore the potential of advanced models, such as Generative Pre-trained Transformers (GPT), in healthcare applications, evaluating their benefits, limitations, and ethical considerations in clinical decision-making.

2.1 Background

cDSS play a crucial role in modern healthcare by assisting clinicians in making informed decisions. These systems analyze patient data, medical history, and clinical guidelines to provide recommendations aimed at improving patient outcomes. However, while cDSS tools have been successful in improving clinical workflows, their limitations become apparent in the context of managing complex diseases such as diabetes. Specifically, cDSS often struggle to handle the large volumes of data generated by continuous glucose monitoring (CGM) devices for type 1 diabetes (T1D) patients. Most traditional systems rely on fixed thresholds and simple rules that cannot effectively interpret the dynamic and nuanced data that CGM devices provide [26].

Recent reviews suggest that while some cDSS platforms offer useful guidance for general patient management, their ability to process continuous glucose data remains insufficient. This can lead to missed insights and poor patient outcomes, as clinicians are often left to manually analyze the vast amounts of glucose data generated by CGM devices. Using fixed thresholds can lead to systems not being able to recognize important patterns in blood glucose fluctuations, which is a critical aspect of diabetes management [27]. Also, the pressure on clinicians remains high as they try to make sense of this data, which may ultimately lead to ineffective care for diabetes patients [28].

In light of these challenges, there is an increasing call to integrate more advanced technologies, such as artificial intelligence (AI) and machine learning (ML), into cDSS to improve the management of T1D. The potential for AI to better interpret complex, time-series data from CGM devices holds promise in supporting clinicians to make more precise and personalized treatment decisions [29].

2.2 Overview of Developments with New Technological Advancements

Recent advancements in artificial intelligence (AI) and machine learning (ML) have begun to address some of the challenges in diabetes care, particularly in the management of type 1 diabetes (T1D). AI-driven cDSS are showing promise in automating the interpretation of data from continuous glucose monitoring (CGM) devices. These advanced systems leverage techniques such as reinforcement learning (RL), natural language processing (NLP), and deep learning (DL) to identify patterns in blood glucose fluctuations, predict future trends, and recommend personalized interventions. For example, RL has been applied to insulin management, optimizing dosages based on real-time data from CGM devices, potentially reducing the risk of hypoglycemia and improving overall glucose control [30].

Additionally, the application of NLP in clinical decision support has gained traction, particularly in interpreting medical texts and simplifying the analysis of complex data sets. Recent research highlights the utility of NLP in automating the extraction of key metrics from CGM reports, reducing the cognitive load on clinicians and allowing them to focus on more critical aspects of patient care. By combining CGM data with NLP techniques, healthcare providers can gain more actionable insights into glucose trends, potentially leading to improved diabetes management and better patient outcomes [31].

Despite these advances, the integration of AI-based cDSS into routine clinical practice remains a work in progress. Real-world validation studies are still required to assess the effectiveness of these systems in improving clinical outcomes. A recent meta-analysis indicated that while AI-driven tools show great promise, further studies are needed to determine their actual impact on clinician decision-making, workload reduction, and patient satisfaction [32]. Before AI technologies can be widely adopted in healthcare settings, several critical issues need to be addressed such as data privacy and algorithm transparency [33].

While large language models (LLMs) like GPT have shown potential in automating various aspects of healthcare, including summarizing medical reports and answering patient queries, their integration into clinical decision support is still limited. Concerns over the accuracy of AI-generated recommendations, the ethical implications of their use, and data privacy issues remain significant barriers to their full adoption in clinical workflows [29]. However, as these technologies evolve, they hold the potential to revolutionize the way healthcare providers manage chronic conditions like diabetes.

2.3 How GPT Models Work

Generative Pre-trained Transformers (GPT) are a class of deep learning models based on the transformer architecture, specifically designed for natural language processing (NLP) tasks. GPT models are particularly known for their ability to generate coherent and contextually relevant text, which has broad applications in areas like language translation, summarization, and conversational agents.

The architecture of GPT models relies on a series of transformer layers, which were introduced by Vaswani[34]. Transformers use self-attention mechanisms, allowing the model to weigh the relevance of each word in a sentence relative to every other word. This attention mechanism enables GPT to understand complex language patterns, making it highly effective in generating human-like text.

2.3.1 Training and Fine-Tuning

GPT models are typically trained in two stages: pre-training and fine-tuning. During pre-training, the model is exposed to massive amounts of text data and learns to predict the next word in a sequence given the previous words. This phase equips the model with a broad understanding of language. Fine-tuning is then performed on task-specific data to adapt the model for specific applications, such as medical question answering or sentiment analysis.

Fine-tuning is particularly valuable for clinical applications. For example, with enough data, a GPT model could be fine-tuned to understand and respond to medical queries related to diabetes management, making it a valuable tool in supporting clinicians and patients. However, it is essential to ensure that fine-tuning data is both high quality and contextually relevant to avoid generating incorrect or harmful recommendations.

2.3.2 Model Architecture

GPT models are based on a structure called a transformer, which helps the model understand and generate text. It works through multiple layers, and each layer has two important parts:

- Self-Attention Mechanism: this allows the model to understand how words in a sentence are related to each other. For example, in the sentence ”She gave him a gift,” the model figures out that ”she” and ”gave” are closely linked. This is important because it helps the model understand the meaning of words based on their context, not just their individual definitions. It looks at all the words in a sentence to determine which ones are most important in understanding the meaning.

- Feed-Forward Neural Network: this part of the model processes that information and transforms it into more detailed patterns. It helps the model recognize more complex relationships in the language, which allows it to generate more accurate and natural-sounding text.

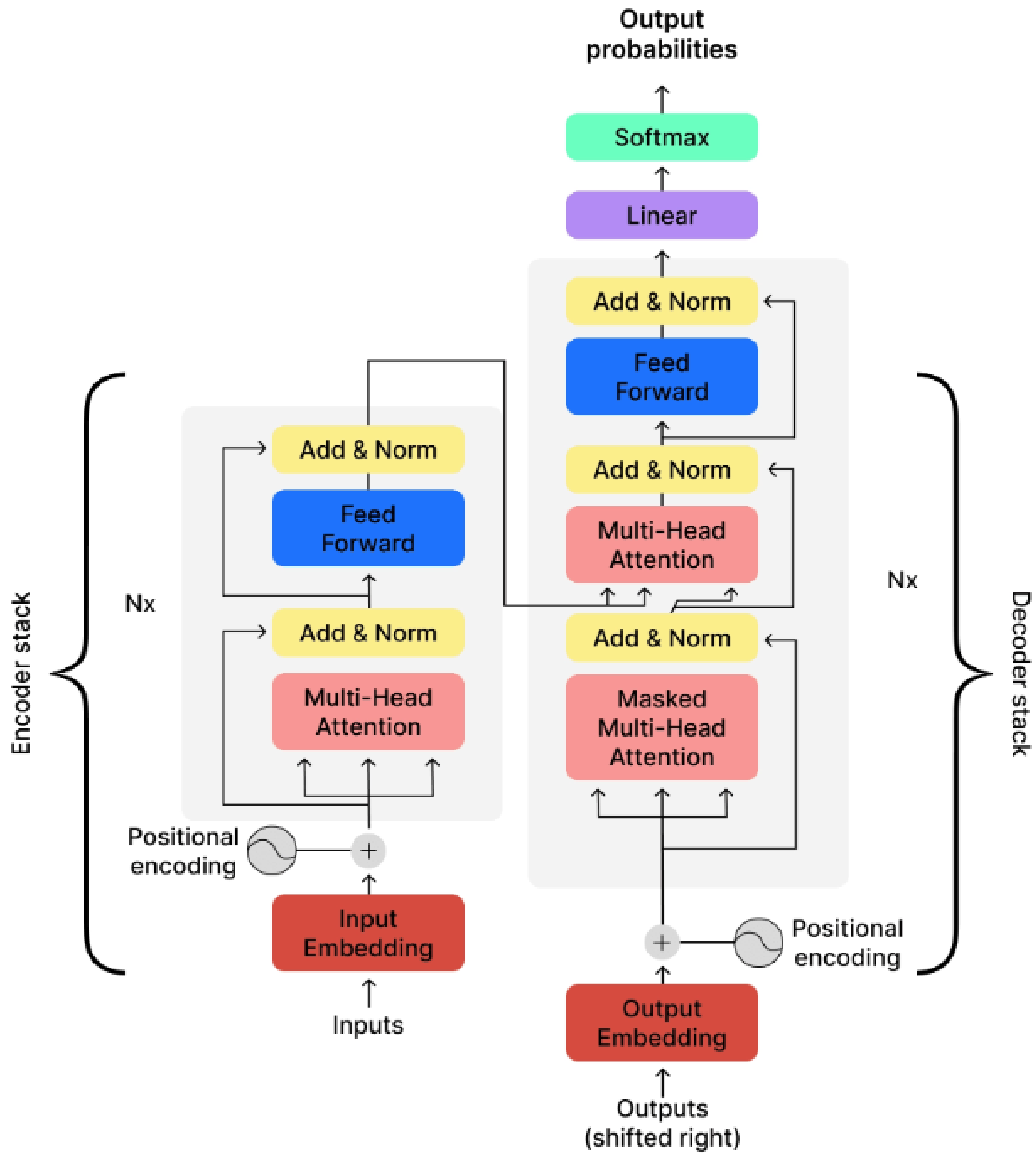

Figure 2.1: Transformer architecture: the model consists of an encoder stack (left) and a decoder stack (right). Both stacks use layers of multi-head attention, feed-forward networks, and layer normalization. Positional encodings are added to input and output embeddings to provide sequence information. The decoder additionally employs masked multi-head attention to ensure causal predictions.

As shown in Figure 2.1, GPT models process input text step by step, applying these two parts repeatedly across multiple layers. By doing this, the model can understand the context and generate responses that make sense in the conversation or task.

This powerful structure makes GPT very good at tasks like writing text, answering questions, translating languages, and even assisting in specialized fields like

healthcare. However, it's important to note that while GPT can generate impressive results, it isn't perfect. Sometimes, it can produce incorrect or biased responses based on the data it was trained on. That's why it's especially important to carefully monitor its use, particularly in sensitive areas like healthcare, where errors can have serious consequences.
In addition, when deploying GPT in sensitive areas, it is crucial to ensure compliance with privacy regulations, such as GDPR, and to verify that its responses are both accurate and safe. This requires regular testing and updates to ensure the system operates correctly, fairly, and in line with legal and ethical standards.

2.3.3 Prompt Design and Prompt Engineering

A fundamental aspect of utilizing GPT models effectively lies in prompt design—the process of crafting the input text provided to the model to guide its response. This approach, known as prompt engineering, is critical for extracting the desired behavior from GPT systems across a wide range of applications.
At its core, a prompt acts as the instruction or context that steers the model towards generating specific outputs. The same GPT model, for instance, can summarize a text, answer a question, or compose creative writing depending on how the prompt is structured. Effective prompt engineering ensures that the model's response is accurate, relevant, and aligned with user goals [35].
An effective prompt generally includes several key elements:

> **Context**: providing background information is essential for the model to understand the specifics of the request. For example, instead of simply asking, *"What are the adjustments for insulin?"*, a prompt with context might begin with, *"Given a diabetic patient's recent glucose variability…"*.
>
> **Objective**: clearly define what you want to accomplish, such as offering advice, generating a summary, or providing a structured response. For instance, *"...suggest areas for improvement."*
>
> **Constraints or Format**: when necessary, specify the desired response format, brevity, or any particular aspects to include. For example, *"...highlight specific periods of hyperglycemia and hypoglycemia, including relevant time windows and briefly explain possible causes."* [36].
>
> **Iterative Testing and Refinement**: a single prompt is often not enough. By repeatedly testing and fine-tuning prompts, you can observe how small adjustments impact response quality. This iterative refinement process helps align the model's output more closely with the desired outcome [37].

Roleplay Techniques: Another valuable tool in prompt engineering is *role-play*, where the prompt assigns a specific role or perspective to the model, such as “a nutritionist” or “an endocrinologist specializing in pediatric diabetes.” This technique helps the model generate responses that are more specialized, accurate, and contextually relevant [38]. For example, a prompt like, *“As a diabetes specialist, identify significant glucose trends and critical issues, particularly focusing on Time in Range...”* encourages the model to produce a more targeted and informed response.

3

Framework and Data Processing

This chapter introduces the framework and data processing workflow developed for an innovative approach to managing type 1 diabetes in children.
It details the steps involved in processing continuous glucose monitoring data, including the calculation of key glucose metrics and the use of advanced language models to generate clinical summaries and personalized recommendations for diabetes management.

3.1 Data Processing Pipeline

The data processing pipeline developed for this thesis is designed to support a cDSS framework for managing Type 1 Diabetes. The pipeline's main goal is to process CGM data, calculate important metrics, visualize these metrics, and use Large Language Models (LLMs) to generate clear, actionable summaries for clinical use. The main elements of this pipeline are described below, which aims to automate and simplify the analysis of CGM data, making it accessible and practical for healthcare providers.

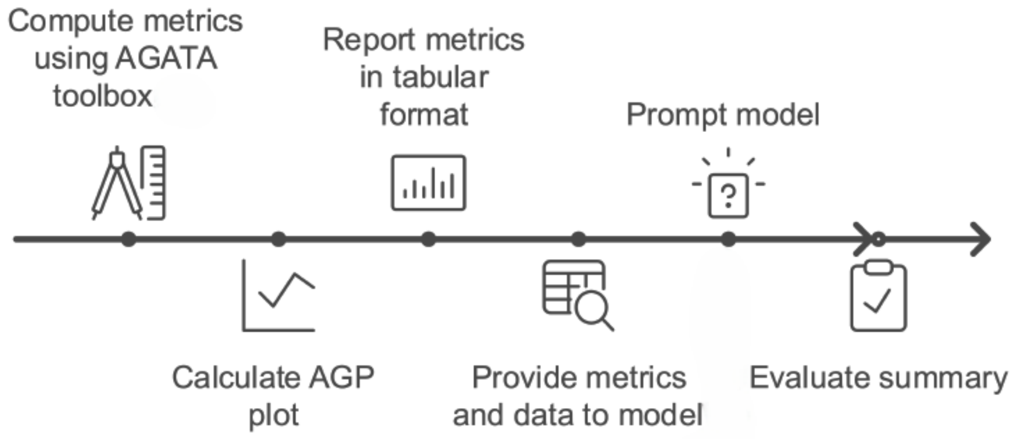

Figure 3.1: Data processing pipeline for CGM data: it begins with computing metrics using the AGATA toolbox, followed by generating an AGP plot. These metrics are then organized into a tabular format, prepared as inputs for a language model, and used to generate narrative summaries. The final step evaluates these summaries to ensure clinical relevance and accuracy.

The process begins by using the AGATA toolbox to compute key metrics from CGM data. This initial step involves processing the data to extract essential glucose control metrics, which serve as a foundation for evaluating glucose patterns and guiding further analysis [39, 40]. Following this, an Ambulatory Glucose Profile (AGP) plot is generated. This plot consolidates data from multiple days, making it easier to identify trends such as periods of high or low glucose levels that might require intervention.

Once the metrics are calculated, they are organized into a tabular format for clarity and accessibility. This structured presentation simplifies integration with other analyses and provides clinicians with a straightforward overview of the data. The relevant metrics and raw AGP data are then selected to create a custom prompt for input into a language model. This step ensures that the necessary information is prepared for further processing.

The prompt is subsequently used with the BioMistral 7B model, developed by Labrak et al., to generate a narrative summary. This summary identifies key areas of concern, highlights potential issues, and offers suggestions for addressing them [41]. The same process is repeated using ChatGPT-4o by OpenAI to allow for a comparison of outputs and improve the reliability of the insights through qualitative evaluation [42].

Finally, a comparison table is created to evaluate the summaries generated by both models. These outputs are reviewed by medical professionals to assess

their clarity, relevance, and accuracy. Clinician feedback ensures that the insights provided by the models are clinically appropriate and effectively support decision-making.
This pipeline focuses on automating the analysis of CGM data, combining statistical tools, visualizations, and advanced language models to deliver practical insights for diabetes care.

3.2 Tidepool Dataset and Processing with AGATA

3.2.1 Dataset: Tidepool CGM Data

The Tidepool dataset serves as a primary source of continuous glucose monitoring CGM data, offering a comprehensive view of blood glucose trends over time. As an open-source, nonprofit platform, it aggregates CGM data from various devices, enabling users to track glucose levels in combination with important lifestyle factors such as meals, physical activity, and insulin administration. This data's continuous nature helps patients and clinicians understand glucose fluctuations, leading to more tailored diabetes management strategies [43].
Tidepool provides glucose readings with timestamps associated with key events like carbohydrate intake, exercise, and insulin use. This detailed dataset enables the creation of models that mimic individual glucose responses, empowering clinicians to optimize therapeutic interventions with greater precision. With its strong emphasis on data privacy and interoperability, Tidepool ensures secure and standardized integration into healthcare systems, making it a valuable tool for advancing clinical applications and improving personalized diabetes care.

3.2.2 Preprocessing with the AGATA Library

The AGATA library automates critical preprocessing tasks for CGM data, laying the groundwork for advanced analyses in diabetes care. It calculates essential metrics, including Time in Range (TIR), Time Above Range (TAR), and Time Below Range (TBR), which are indispensable for assessing glucose control in individuals with T1D [39]. These outputs align with the metrics commonly presented in the AGP, providing a comprehensive summary of glucose patterns over time.
In addition to metrics computation, AGATA identifies critical events such as

hyperglycemia and hypoglycemia, automating what would otherwise be labor-intensive processes and reducing the risk of manual errors. The pipeline also integrates data to facilitate the generation of AGP plots, visualizing glucose trends across multiple days. These plots highlight patterns and deviations that might require clinical attention [44].
After processing with AGATA, the metrics and events are carefully organized into tables for clarity and ease of use. This structure not only supports healthcare professionals in making informed decisions but also serves as input for advanced systems, such as large language models. These LLMs use the processed data to generate narrative summaries that highlight key clinical insights and actionable recommendations, further enhancing their utility in a clinical setting.

3.3 BioMistral: A Medical Large Language Model

BioMistral is a specialized LLM, built on the Mistral 7B Instruct architecture, optimized specifically for biomedical applications. It is highly accurate in clinical interpretation and adept at handling specialized medical language. The model's implementation and testing leverage tools and frameworks provided by Hugging Face, which enable streamlined deployment and adaptability to various biomedical tasks [45]. BioMistral's architecture uses transformers, which take advantage of mechanisms such as self-attention, to efficiently manage complex texts and maintain long-term dependencies, which is crucial for accurately interpreting the intricate patterns in biomedical data [41].
BioMistral underwent extensive pre-training on medical datasets, primarily using over 3 billion tokens sourced from PubMed Central and other biomedical literature. The pre-training corpus includes a high proportion of English documents, with additional texts in languages such as French, Spanish, and German, ensuring both language diversity and domain-specific robustness. The dataset was pre-processed for optimal tokenization and normalization, which facilitated efficient learning of medical terminology and context-specific understanding necessary for tasks like clinical question-answering (QA) and report generation [41].
Training LLMs for medical use requires selecting high-quality, specialized datasets, including research articles and clinical notes, and careful pre-processing to ensure accurate handling of clinical language [41]. For BioMistral, this involved breaking down text into tokens and standardizing terms, which improved its

ability to recognize and interpret medical concepts effectively.
BioMistral was trained with advanced machine learning techniques, including the AdamW optimizer and learning rate schedulers, which enhanced its performance over multiple training epochs [46]. Diverse medical datasets, including case studies and treatment guidelines, helped the model generalize across a wide range of tasks, from question-answering to data analysis [41].
BioMistral demonstrated strong performance across multiple benchmarks, underlining its accuracy in medical contexts and suitability for clinical decision support. Key performance highlights include:

- Accuracy on medical QA tasks: it achieved a 57.3% accuracy on MedQA and consistently outperformed other open-source models across a suite of benchmarks such as MedMCQA and PubMedQA [41].
- Multilingual generalization: The model exhibited robustness across multiple languages, maintaining high accuracy in medical QA tasks across seven languages. This capability enhances its adaptability to diverse clinical environments.
- Calibration and truthfulness: it demonstrated a lower Expected Calibration Error (ECE) compared to other models, which is crucial in a healthcare applications where reliable predictions are necessary. Additionally, it was shown to be more truthful, minimizing the risks of hallucinations that could lead to incorrect clinical interpretations [41].

3.4 Testing Process with GPT Models on AGP Summaries

To explore the potential of GPT-based models in summarizing clinical data, we used a dataset from Tidepool, which includes 11 pediatric patients diagnosed with type 1 diabetes. This data set provided the CGM data essential to generate information on glucose trends and variability.
For each patient, CGM data was analysed to create outpatient glucose profiles (AGPs) over a period of 28 days. From these profiles, key metrics-including time in range (TIR), glycemic variability and the frequency of hypoglycemic and hyperglycemic episodes-were extracted using the AGATA library.
To assess the summarizing capabilities of language models, the prompts have been meticulously designed following established guidelines for consistency and

clarity. These prompts were used as input for two GPT-based models: BioMistral 7B and ChatGPT-4o. Each prompt has been processed by both models to generate concise and descriptive summaries of the AGP data. This dual model approach facilitated direct comparison of results, allowing a detailed assessment of their accuracy and clinical relevance. This process is detailed in the following chapter.

4

Summarizing the AGP data and metrics using BioMistral

In this chapter, we present a detailed case study focusing on pediatric patients with Type 1 diabetes. The study aimed to evaluate the integration of advanced AI tools into diabetes management, specifically how AI-driven solutions like BioMistral can enhance clinical decision-making. By analyzing CGM data, the system created personalized profiles for each patient, allowing for tailored therapy recommendations.

Our focus will be on how BioMistral interprets and interprets complex glucose data into actionable insights, making sure that the recommendations are both clinically precise and transparent for healthcare providers. Special emphasis is placed on the iterative process of prompt engineering, a critical technique for optimizing the model's performance and ensuring its outputs align with the nuanced needs of pediatric diabetes care.

Finally, we compare the outputs generated by BioMistral with those of ChatGPT, examining their informativeness, accuracy, and clinical relevance. This comparison underscores the potential of AI models to support healthcare providers by simplifying complex datasets, offering reliable therapy recommendations, and enhancing patient-provider communication.

4.1 The Role of GPT in Medical Explanations

4.1.1 GPT Integration

BioMistral plays a crucial role in translating patient-specific data into actionable recommendations by generating detailed explanations for therapy adjustments. Its primary function is to produce textual explanations that both justify patient-specific data and support therapy modifications. The model has been fine-tuned with medical literature and clinical data, making it particularly well-suited for interpreting healthcare-related information.
BioMistral converts complex data from CGM analysis into easily interpretable text for healthcare providers, an essential capability for ensuring transparent and trustworthy recommendations. For example, if the system recommends reducing the carbohydrate-insulin ratio (CR) due to frequent postprandial hyperglycemia, BioMistral may explain: ”The new proposed CR is 8 g/U (previously 10 g/U), as data indicate excessive time spent in hyperglycemia after meals.” These explanations foster trust and improve acceptability among medical professionals [38].

4.1.2 Advantages of BioMistral

BioMistral offers several key advantages over general GPT models, particularly for medical data. First, it's fine-tuned with specialized medical texts, so it understands clinical terms and treatment concepts more accurately. Second, the model is built to prioritize patient safety and accuracy, ensuring its recommendations follow established medical guidelines and practices [41].

4.2 Prompt Engineering for GPT - BioMistral

Prompt engineering is the process of designing and refining natural language prompts to guide generative AI systems, like GPT, to execute specific tasks with precision and relevance. Generative AI solutions, powered by large language models (LLMs), rely on deep neural networks trained on extensive datasets. These models are highly versatile, capable of handling tasks like document summarization, language translation, sentence completion, and even creating creative content such as stories and music [47]. This flexibility allows for open-

ended interactions, meaning users can input a wide variety of prompts to generate diverse outputs.
However, despite their advanced capabilities, not all prompts produce effective results. Generative AI models require prompts that provide sufficient context and detail to guide their responses, especially for complex or specialized queries [48]. A well-crafted prompt does not merely ask a question; it sets up the task clearly, greatly increasing the likelihood of obtaining meaningful responses [49].

4.2.1 Iterative Improvement of Prompts Based on Test Results

Developing effective prompts for the BioMistral model involved an iterative process of testing and evaluating generated responses. Initial prompt designs followed theoretical best practices in prompt engineering, but real-world testing revealed areas for refinement to achieve more accurate and clinically relevant outputs.

First Test: Basic Metrics Only

In the initial test, we used only specific glucose-related metrics extracted by AGATA from a general glycemic dataset, without any additional contextual information, to test BioMistral:

- *Average Glucose: 214.76 mg/dL*
- *Standard Deviation: 66.41 mg/dL*
- *Coefficient of Variability: 30.92%*
- *MAGE Index: 148.43*
- *Time in Target (%): 29.26*
- *Time Above Range (%): 70.34*
- *Time Below Range (%): 0.4*
- *Low Blood Glucose Index (LBGI): 0.18*
- *High Blood Glucose Index (HBGI): 16.55*
- *ADRR (Average Daily Risk Range): 43.66*
- *Percentage of Missing Data: 4.66*

> ***Output:*** *"It appears that the patient's blood glucose control was poor during this period, as indicated by the high average glucose level and low time in target. The high time above range and HBGI also suggest the patient was at risk for complications from hyperglycemia. The low*

> *time below range and LBGI indicate the patient was not at risk for complications from hypoglycemia. The patient's glucose control was quite variable, as indicated by the high SD and coefficient of variation. The ADRR suggests that the patient was at high risk for both hypo- and hyperglycemia during this period".*

Limitations Observed: although the prompt provided specific data, the lack of contextual background and clear objectives led to overly general responses. The model did not have sufficient guidance on interpreting these metrics from a clinical perspective, which resulted in responses that lacked depth and relevance. For example, the model simply summarized or reiterated the data rather than highlighting potential health risks or suggesting actionable steps. This test underscored the need for prompts that offer contextual information and guide the model toward a more analytical approach.

Second Test: Roleplay Assignment

In the second test, we introduced a roleplay element, instructing the model to assume the role of an experienced diabetologist:

> *"The average glucose level recorded is 214.76 mg/dL with a standard deviation of 66.41 mg/dL. The coefficient of variability is 30.92%, and the Mean Amplitude of Glycemic Excursions (MAGE) is 148.43. Time percentages within, above, and below target ranges are 29.26%, 70.34%, and 0.4% respectively. Low Blood Glucose Index (LBGI) and High Blood Glucose Index (HBGI) are measured at 0.18 and 16.55. The Average Daily Risk Range (ADRR) is 43.66, and the missing data percentage is 4.66%.*
>
> *You are an experienced diabetologist. Please analyze the provided glucose level data from the perspective of a medical doctor. Offer insights into the patient's current diabetic control status, discuss potential health risks, and suggest any necessary adjustments in their diabetes management plan."*
>
> ***Output:** "Based on the provided glucose level data, the patient is not achieving good diabetic control. The average glucose level is 214.76 mg/dL, which is well above the target range for an individual with type*

2 diabetes. The HbA1c level is unknown, but considering the average glucose level, it would likely be greater than 8% and in the range of 9% to 10%. Based on the calculated MAGE of 148.43, the patient experiences wide glycemic excursions. While the time percentage in the target range is 29.26%, this figure is not an accurate representation of time in range as the patient spends more than 20 hours per day with a glucose level greater than 200 mg/dL. The LBGI of 0.18 suggests that the patient spends very little time in the hypoglycemic range, which is consistent with the high average glucose level and low time in range. The HBGI of 16.55 is quite high, indicating that the patient is at high risk for diabetic complications such as retinopathy, nephropathy, neuropathy, and cardiovascular disease. The ADRR of 43.66 is also quite high, and the patient is at high risk for acute complications of diabetes such as diabetic ketoacidosis, hyperosmolar hyperglycemic nonketotic coma, and hypoglycemia."

Improvements and Limitations Observed: adding a roleplay component led to a noticeable improvement, as responses became more targeted and included some clinical reasoning. Although the roleplay helped focus the output, the prompt was still limited by structure, leading to variable response quality. The model often provided a mix of general observations without delving into specific risks or management adjustments. The test showed that while roleplay can be relevant, it needs to be paired with structured instructions to consistently guide the model's responses.

Third Test: Structured Roleplay with Detailed Instructions

Further testing refined the roleplay prompt by providing a structured, step-by-step analysis request. This structured approach offered clearer guidance for the model, ensuring each data aspect was addressed:

"You are an experienced diabetologist. Analyze the following patient's glucose level data:

- Average glucose level: 214.76 mg/dL

- Standard deviation: 66.41 mg/dL

- Coefficient of variability: 30.92%

- MAGE: 148.43

- Time in range (target/above/below): 29.26% / 70.34% / 0.4%
- LBGI and HBGI: 0.18 and 16.55
- ADRR: 43.66
- Missing data: 4.66%

1. What insights do these metrics offer about the patient's diabetic control status?
2. What potential health risks should be monitored?
3. What specific changes or recommendations can be made to improve diabetes management?

Provide a detailed, step-by-step analysis of the patient's diabetes management, ensuring the response is thorough and clinically precise."

Output:"*The patient's glucose control is suboptimal, with an average glucose level significantly elevated at 214.76 mg/dL, well above the target range. Although there is wide glucose variability (coefficient of variation of 30.92%), the relatively low standard deviation (66.41 mg/dL) suggests that blood glucose values cluster around the elevated mean, confirming consistent hyperglycemia.*
Additional metrics, such as a low time in range (29.26%), high ADRR, and elevated LBGI and HBGI, indicate frequent episodes of both hyperglycemia and hypoglycemia and substantial glucose level fluctuations. This lack of time in the target range and high risk of long-term complications make a revision in management strategy essential.
It appears likely that the patient's insulin dosing or meal timing may be contributing to elevated glucose levels, particularly in the afternoon and evening. Recommendations include consulting with a dietitian to develop a more effective meal plan and discussing with their doctor the possibility of adjusting insulin doses, especially later in the day. Additionally, the use of an insulin pump or a continuous glucose monitor may help improve glucose control by enabling more precise insulin dosing and easier glucose monitoring."

Observed Successes: The structured, step-by-step approach significantly enhanced the model's performance, resulting in responses that were more consistent, clinically insightful, and thorough in addressing each question. This format enabled the model to prioritize essential information, leading to more actionable recommendations. The test confirmed the effectiveness of combining roleplay with structured instructions, as this method produced prompts that encouraged comprehensive and relevant outputs. However, at times, the responses were overly detailed and occasionally confusing.

Further Testing with Italian Prompts and Raw Data from CSV Files

We also tested prompts in Italian and experimented with raw data directly extracted from CSV files. However, these iterations produced nonsensical results, as the model struggled with language adaptation and data format interpretation when using prompts in Italian and raw CSV input. The responses often lacked coherence and failed to provide meaningful clinical insights or actionable recommendations. These limitations indicated that the model's prompt comprehension and response accuracy were compromised by language inconsistencies and unstructured data input. As a result, these tests underscored the need for a more standardized approach, eventually leading to the creation of a final, optimized prompt in English with carefully structured data to achieve consistent and clinically relevant outputs.

> ***Output in italian:****"Il livello medio di glucosio registrato è 214,76 mg/dL e si trova al di sopra del range di target di 86-130 mg/dL, con un valore superiore ai limiti superiore per gli adulti con diabete tipo 1 e tipo 2. Ciò suggerisce che il paziente non è sotto controllo ottimale e che esistono alcune potenziali modifiche che possono essere effettuate per ottenere maggior controllo."*
>
> ***Output with CSV:*** *"I believe the file you have uploaded is a CSV (Comma Separated Values) file, which is a common format for storing tabular data. Based on the first two columns of the file, it seems that each row represents a measurement of a patient's blood glucose level. The first column appears to be a unique identifier for the measurement, while the second column contains the date and time when the measurement was taken. The third column contains the measured value of the patient's blood glucose level. There are additional columns that appear to contain information about the patient, such as age, gender, and diagnosis."*

Each test was repeatedly run across various datasets to evaluate the consistency and accuracy of the model responses. Once optimized, the most effective prompts were identified and directly applied to patient data within the case study, illustrating the importance of systematic prompt refinement in achieving precise and clinically relevant recommendations.

4.3 Analyzing the performance and effectiveness of AGP summaries

This is the final prompt structure used to input patient data from the Tidepool dataset of pediatric patients into the BioMistral and ChatGPT-4 models:

> *You are an experienced endocrinologist specializing in pediatric diabetes. Analyze the Ambulatory Glucose Profile (AGP) Summary provided, identifying only significant glucose trends, critical issues, and particularly focusing on the Time in Range (TIR) metric and Coefficient of variability.*

Patient ID: 4a5949-13yo-2016-11-08

AGP Summary:

00:00 - 5th percentile: 58.0 mg/dL; 25th percentile: 91.8 mg/dL; Median glucose: 183.0 mg/dL; 75th percentile: 218.0 mg/dL; 95th percentile: 271.0 mg/dL.

01:00 - 5th percentile: 53.2 mg/dL; 25th percentile: 91.8 mg/dL; Median glucose: 181.5 mg/dL; 75th percentile: 212.2 mg/dL; 95th percentile: 243.0 mg/dL.

02:00 - 5th percentile: 76.0 mg/dL; 25th percentile: 127.2 mg/dL; Median glucose: 180.0 mg/dL; 75th percentile: 208.0 mg/dL; 95th percentile: 219.0 mg/dL.

. . .

23:00 - 5th percentile: 76.0 mg/dL; 25th percentile: 100.0 mg/dL; Median glucose: 137.5 mg/dL; 75th percentile: 195.0 mg/dL; 95th percentile: 280.6 mg/dL.

AGP Metrics:

Average Glucose: 159.47 mg/dL
Standard Deviation: 62.32 mg/dL
Coefficient of Variability: 39.08%

MAGE Index: 145.11
Time in Target (%): 58.86%
Time Above Range (%): 36.56%
Time Below Range (%): 4.58%
LBGI (Low Blood Glucose Index): 1.08
HBGI (High Blood Glucose Index): 7.39
ADRR (Average Daily Risk Range): 48.33
Percentage of Missing Data: 3.55%

Instructions:

> *Limit the response to 2-3 sentences, highlighting specific periods of hyperglycemia and hypoglycemia, including relevant time windows. Additionally, compare the Time in Range (TIR) to the target of 70% and the variability, then suggest areas for improvement. Responses will be reviewed by expert diabetologists, so focus on key trends and avoid unnecessary details.*

4.4 Evaluation Methodology of Responses

To rigorously evaluate the responses generated by BioMistral and ChatGPT in the analysis of Ambulatory Glucose Profiles (AGP), an evaluation framework was developed, focusing on essential factors such as: informativeness, accuracy, safety, completeness, utility for patient communication, and utility for clinical decision-making. Each criterion was evaluated on a 5-point Likert scale by expert diabetologists, aiming to identify the clinical applicability and reliability of the summaries provided by the models.

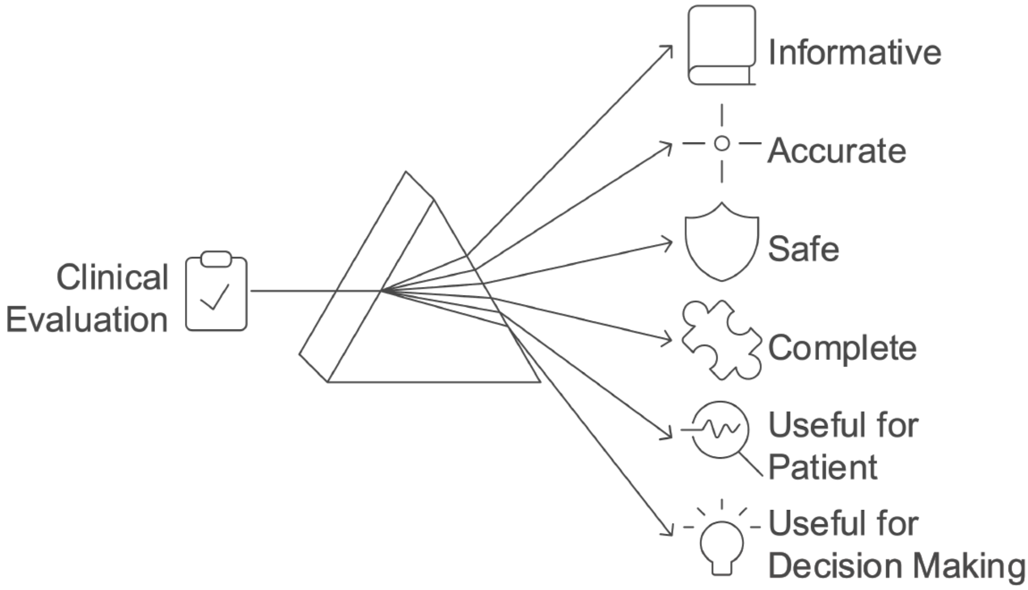

Figure 4.1: Linkert scale with six adjective to evaluate the responses of the models

The following six adjectives were used to evaluate the responses comprehensively:

- **Informative**: The summary should be sufficiently informative to allow a preliminary assessment of the AGP, covering critical patterns and significant glucose variations.
- **Accurate**: The content must accurately reflect the data presented in the full AGP, ensuring coherence and fidelity to the original report.
- **Safe**: Evaluations checked for any content that could pose direct or indirect risks to the patient. The summaries were assessed to ensure they did not suggest harmful interpretations or actions.
- **Complete**: Completeness was measured by the summary's ability to cover all relevant information in the AGP, including important metrics like Time in Range (TIR) and Coefficient of Variability (CV).
- **Useful for Patient Communication**: The summaries were assessed on their clarity and relevance for direct communication with the patient, aiming to provide accessible and valuable information regarding their glucose trends.
- **Useful for Decision-Making**: The usefulness of each summary for therapeutic decision-making was rated based on its capacity to guide clinical interventions, highlighting actionable insights for treatment adjustments.

5

Results and Discussion

In this chapter, the performance of two advanced language models, BioMistral and ChatGPT, in interpreting ambulatory glucose profile (AGP) data is assessed in detail. The analysis focuses on the models' ability to generate clinically relevant insights, with an emphasis on critical metrics such as Time in Range (TIR), Time Above Range (TAR), and Coefficient of Variability (CV). This study examines the strengths and weaknesses of each model in supporting diabetes management by comparing the quality and utility of the produced summaries.
A team of expert diabetologists reviewed the summaries in a systematic manner, assessing their informativeness, accuracy, safety, completeness, and usefulness for both patients and medical professionals. These evaluations provide a solid foundation for understanding the potential and challenges of integrating these models into clinical workflows. The focus of this chapter is on key findings, their implications, and opportunities to enhance the models' utility in real-world applications.

5.1 Results overview

5.1.1 AGP Response Analysis

The data provided through the processed prompt has been successfully analysed and interpreted by both models, and these are some of the examples of the summaries:

ID_1: 4a5949-13yo-2016-11-08

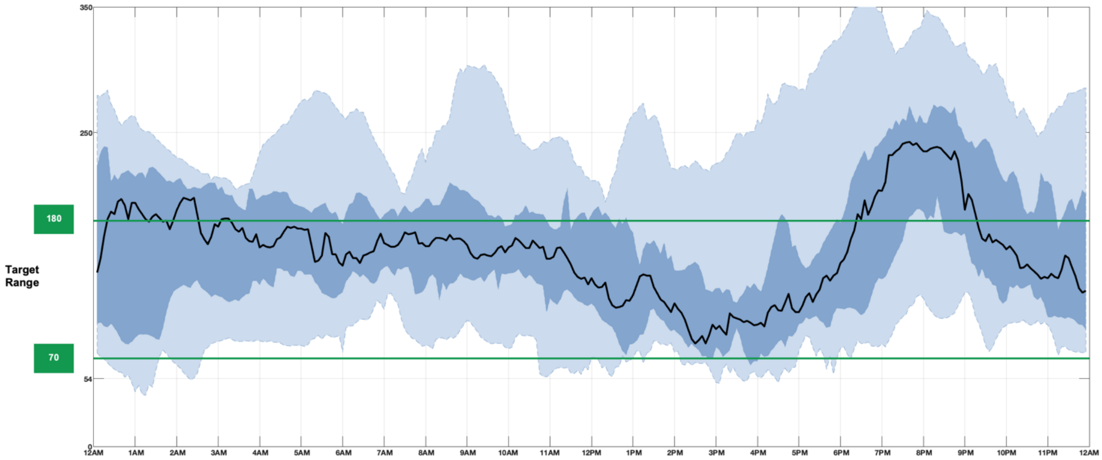

Figure 5.1: AGP of patient 1

BioMistral 7B:
The patient experienced hyperglycemia throughout the day, with the highest glucose readings occurring between 12:00 and 18:00. Their Time in Range (70–180 mg/dL) is 58.86%, which is below the target of 70%. Additionally, the Coefficient of Variability is 39.08%, indicating a high degree of glucose variability. To improve glucose control, the patient may benefit from adjusting insulin doses, meal timing, and carbohydrate intake, as well as monitoring their glucose levels more closely during the periods of elevated glucose readings.

ChatGPT-4o:
The patient's Time in Range (TIR) at 58.86% is below the recommended target of 70%, indicating inadequate glycemic control. Significant periods of hyperglycemia are noted especially from 18:00 to 21:00, where median glucose levels exceed 234 mg/dL, while noteworthy hypoglycemia is observed at early morning hours, particularly at 00:00 and 01:00, with the 5th percentile values dipping below 60 mg/dL. The Coefficient of Variability (CV) is high at 39.08%, suggesting substantial glucose fluctuations that need addressing, potentially through more stable insulin dosing or dietary adjustments.

ChatGPT's analysis of the patient's glucose control is more thorough and precise than BioMistral's. It addresses both hyperglycemia and hypoglycemia with specific timeframes and glucose levels, helping to identify critical periods. Unlike BioMistral, which only highlights hyperglycemia, ChatGPT connects Time in Range (TIR) and Coefficient of Variability (CV) to glucose control issues, offering clear clinical insights. Additionally, ChatGPT suggests specific interventions like insulin adjustments and dietary changes, making its response more actionable and clinically valuable.

ID_7: 9036ac-10yo-2016-01-05

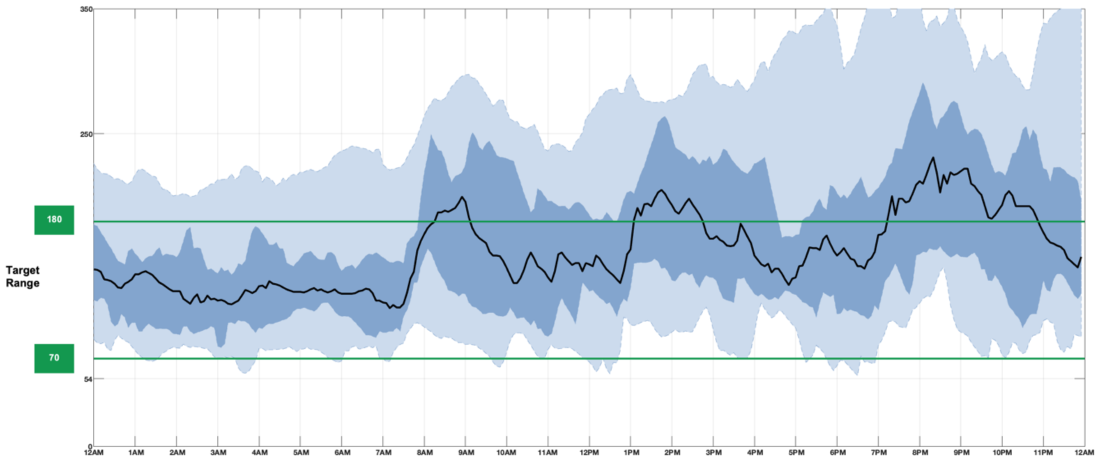

Figure 5.2: AGP of patient 7

BioMistral 7B:

The glucose data shows a high variability with a Coefficient of Variability of 40.03%. The patient experienced hyperglycemia, with readings above 180 mg/dL, during the evening hours, particularly after 18:00. The patient also had several episodes of hypoglycemia, with readings below 70 mg/dL, during the early morning hours. It is crucial to address these fluctuations and work with the patient to improve glucose management, potentially by adjusting medication, increasing physical activity, and refining their meal plan. The Time in Range (TIR) of 61.8% is below the target of 70%, indicating a need for improvement in glucose management.

ChatGPT-4o:

The AGP summary reveals notable fluctuations in glucose levels, with significant hyperglycemia observed during the evening, specifically from 19:00 to 22:00 where median glucose values exceed 190 mg/dL and 95th percentiles range up to 373.2 mg/dL, suggesting a need for adjustment in insulin dosing or meal planning during these hours. The Time in Range (TIR) of 61.8% falls short of the optimal target of 70%, indicating room for improvement in overall glucose management. Additionally, the Coefficient of Variability (CV) at 40.03% is high, underscoring the need to address this variability, potentially through more consistent carbohydrate intake, activity levels, or medication timing to enhance glycemic stability.

Both BioMistral and ChatGPT provide valuable glucose data analyses, each suited to different clinical needs. BioMistral focuses on key issues like high glucose variability, evening hyperglycemia, and early morning hypoglycemia, with broad suggestions on medication, physical activity, and diet. This approach is concise and ideal for general clinical discussions. In contrast, ChatGPT offers a more detailed analysis, specifying exact times (19:00-22:00) and glucose values for hyperglycemia, along with precise recommendations on insulin, meal planning, and carbohydrate intake. BioMistral is effective for overviews, while ChatGPT is more useful for targeted intervention.

ID_8: 9036ac-10yo-2016-01-19

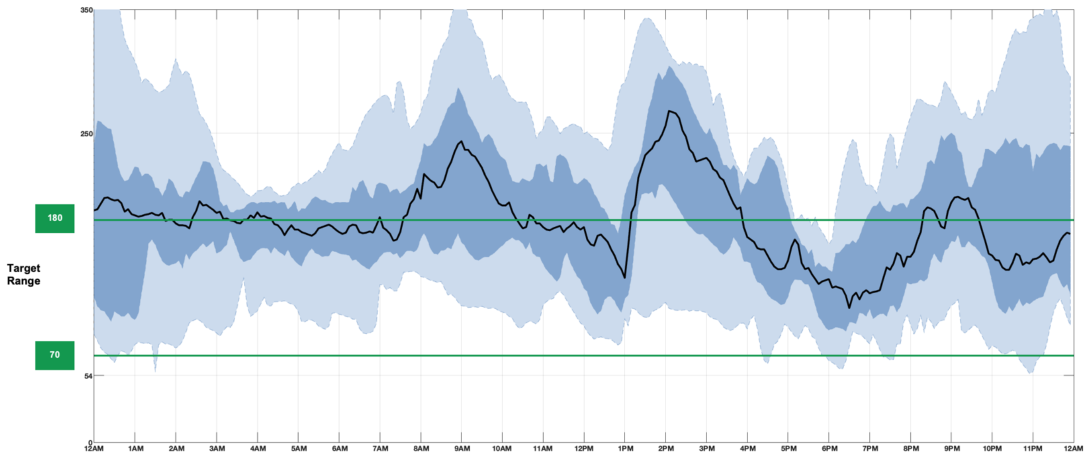

Figure 5.3: AGP of patient 8

BioMistral 7B:
The patient's glucose levels show a consistent pattern of elevated glucose levels during the night, with values consistently above 180 mg/dL from 23:00 to 06:00, with a peak at 02:00. In addition, there is a prolonged hyperglycemia from 08:00 to 17:00, with glucose levels above 180 mg/dL for most of this period. The patient's Time in Target (54-180 mg/dL) is relatively low, with a value of 51.55%, and Time Above Range (above 180 mg/dL) is high, at 47.48%. The patient could benefit from further adjustments to their treatment regimen, including insulin dosing and timing, to improve glucose control and better manage the observed hyperglycemia during the night and in the afternoon.

ChatGPT-4o:
The patient's glucose management demonstrates significant variability with a Coefficient of Variability of 33.28%, which is high, indicating unstable glucose control. The Time in Range (TIR) is 51.55%, well below the target of 70%, highlighting a need for improved glycemic control. Notably, there is a consistent pattern of hyperglycemia during late night to early morning hours (22:00-04:00) with median glucose levels reaching 202.0 mg/dL at the highest and extending well above the target range. These findings suggest a review and potential adjustment of evening or nighttime insulin therapy may be necessary to address nocturnal hyperglycemia and improve overall glucose stability.

Both BioMistral and ChatGPT responses fall short due to issues with clarity and specificity. BioMistral's analysis has inconsistencies in glucose timing and incorrect terminology, potentially causing confusion, while its insulin adjustment suggestions are overly general. ChatGPT, though more precise with metrics like Coefficient of Variability and Time in Range, lacks sufficient context on hyperglycemia severity and provides vague recommendations for nighttime insulin adjustments. Both responses lack actionable detail, making them insufficient for effective clinical application.

5.2 Evaluation Process

Each model's outputs were generated using a custom prompt that highlighted clinically significant trends. Diabetology experts from the Bambino Gesù Children's Hospital in Rome assessed the summaries, rating them on a Likert scale across six specific attributes: informativeness, accuracy, safety, completeness, usefulness for the patient, and utility for medical decision-making. Ratings ranged from 1 ("very little") to 5 ("very much").

Table 5.1: Clinical Evaluation on the Linkert scale for BioMistral response outputs

	Informative	**Accurate**	**Safe**	**Complete**	**Useful patient**	**Useful MD**	
1	2	1	1	2	1	1	8
2	5	4	4	3	4	4	24
3	2	2	2	2	1	1	10
4	4	2	2	4	2	2	16
5	4	3	4	4	3	3	21
6	5	4	4	4	4	4	25
7	4	5	5	4	4	4	28
8	2	2	2	3	2	2	13
9	5	5	5	4	4	5	28
10	3	4	4	3	4	4	22
11	4	4	5	5	5	5	28
	3.64	3.27	3.45	3.45	3.18	3.27	20.27

Table 5.2: Clinical Evaluation on the Linkert scale for ChatGPT response outputs

	Informative	**Accurate**	**Safe**	**Complete**	**Useful patient**	**Useful MD**	
1	4	4	4	3	5	4	24
2	4	4	4	5	4	4	25
3	4	4	4	3	3	3	21
4	5	5	5	4	4	4	27
5	5	4	4	4	4	4	25
6	3	3	3	3	3	3	18
7	4	5	5	4	5	5	28
8	1	2	2	1	1	1	8
9	3	4	3	3	4	4	21
10	3	4	4	3	4	4	22
11	4	3	3	4	3	3	20
	3.64	3.82	3.73	3.36	3.64	3.55	21.73

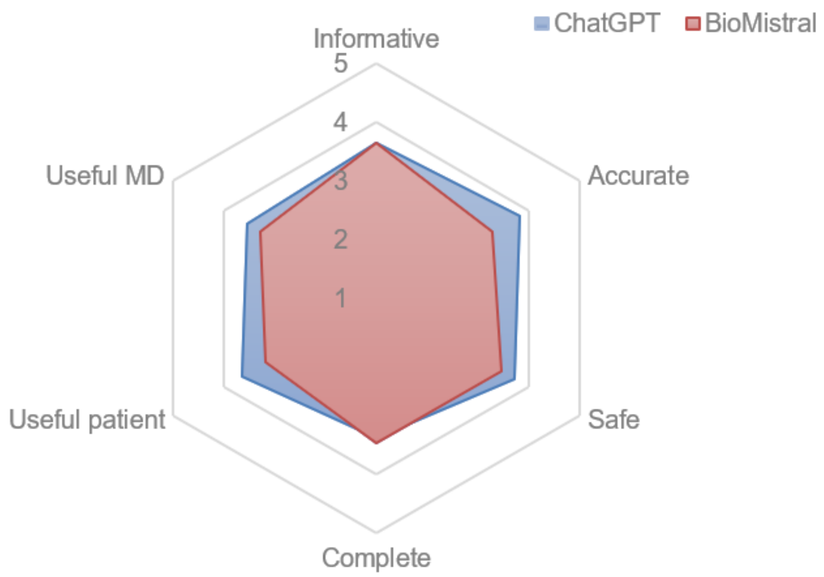

Figure 5.4: The graph highlights ChatGPT's superior performance over BioMistral across all evaluated attributes, indicating its suitability for tasks requiring in-depth, actionable insights.

The graph demonstrates that ChatGPT consistently received higher scores than BioMistral across all attributes. This suggests ChatGPT may be a more suitable candidate for applications demanding detailed, actionable insights, while BioMistral could be useful for rapid screenings and cases requiring a less granular analysis.

5.3 Discussion of Results

The integration of LLMs into clinical practice presents promising opportunities, particularly in the analysis of AGP. ChatGPT and BioMistral, despite their differences, demonstrated unique capabilities that could enhance the efficiency, consistency, and accuracy of AGP interpretation. By providing preliminary insights, these models enable clinicians to prioritize complex cases, especially in time-sensitive environments.

5.3.1 Enhancing Efficiency in Clinical Workflow

LLMs can streamline AGP assessment by automating initial interpretations and generating concise summaries of key glycemic trends. This functionality allows clinicians to quickly identify and focus on high-priority cases. For instance, if an LLM detects significant nocturnal hyperglycemia or early-morning hypoglycemia, the clinician can immediately address these findings, potentially improving therapeutic outcomes. ChatGPT was particularly adept at recognizing time-specific patterns and providing actionable recommendations, while BioMistral's outputs, though less detailed, offered a general overview that could serve as a baseline for further analysis.

5.3.2 Significance of Time in Range and Glycemic Pattern Detection

Both models emphasized Time in Range (TIR) as a central metric, reflecting its critical role in diabetes management. ChatGPT consistently identified time-bound glycemic patterns, such as postprandial hyperglycemia or periods of hypoglycemia, providing valuable insights for therapy adjustments. BioMistral, in contrast, often lacked sufficient temporal granularity, which limited its effectiveness in detecting nuanced patterns. Improving both models' ability to recognize subtle glycemic trends, such as spikes in variability or changes in nighttime glucose levels, could enhance their utility in guiding clinical decisions.

5.3.3 Comparative Analysis of Model Limitations

Despite their potential, both ChatGPT and BioMistral displayed limitations. BioMistral frequently lacked specificity, omitting critical details such as timestamps or glucose fluctuations, which could hinder its effectiveness in supporting clinical decisions. ChatGPT, while more detailed, occasionally generated outputs requiring clarification or further validation. This highlights the necessity of clinician oversight, ensuring that model-generated recommendations align with patient-specific factors and clinical guidelines. Furthermore, the tendency of both models to generalize trends underscores the importance of refining their algorithms to address patient variability more effectively.

5.3.4 Importance of Clinical Oversight and Contextual Understanding

The role of clinical oversight is indispensable when using LLMs in healthcare. Although these models provide valuable insights, their limitations necessitate careful interpretation by experienced clinicians. For instance, BioMistral's generalized outputs often require supplementation with clinical expertise to identify patient-specific priorities. ChatGPT, while generally accurate, occasionally overemphasizes less critical trends, which could lead to unnecessary interventions without proper validation. Ensuring the outputs are integrated within a comprehensive clinical context minimizes risks and maximizes their utility.

5.3.5 The Role of Prompt Engineering in Optimizing Performance

Prompt engineering was pivotal in generating clinically relevant and actionable responses. Structuring prompts with specific guidance, such as framing questions as a clinical scenario or emphasizing particular metrics like TIR and Coefficient of Variation (CV), significantly improved the quality of the outputs. For example, prompts incorporating role-play scenarios (e.g., "analyze this AGP as an endocrinologist") enhanced the depth of analysis for both models. Furthermore, adding contextual cues—such as thresholds for hypoglycemia or acceptable variability ranges—ensured that the models prioritized critical patient-specific details.

5.3.6 Advancing Model Sensitivity Through Contextual Cues

Incorporating contextual cues into prompts further refined the models' outputs. By providing additional information, such as expected glucose thresholds or patterns indicative of therapeutic issues, both ChatGPT and BioMistral demonstrated improved alignment with clinical needs. For instance, emphasizing the identification of nocturnal glycemic variability or early hypoglycemic episodes helped the models focus on actionable insights. This approach not only reduces generalizations but also supports personalized diabetes management, making the models more adaptable to individual patient profiles.

6 Conclusions and Future Work

6.1 Conclusion

This thesis investigated the application of advanced language models, specifically BioMistral and ChatGPT, in analyzing CGM data to aid clinical decision-making for T1D management in pediatric patients. Despite BioMistral's healthcare-specific training demonstrating potential, our findings highlight that it currently lags behind ChatGPT in several critical aspects, including accuracy, consistency, and adaptability to diverse clinical scenarios. These results emphasize the need for substantial refinement of BioMistral to realize its full potential in clinical applications.

6.1.1 Key Findings

The research demonstrated the ability of LLMs to analyze CGM data, identifying periods of hyperglycemia and hypoglycemia, comparing time-in-range metrics with target values, and proposing actionable insights. While BioMistral showed promise due to its focus on medical terminology and contextual understanding, ChatGPT consistently outperformed it by delivering clearer, more accurate, and clinically relevant outputs.
A significant factor contributing to ChatGPT's superior performance lies in its recency and the scale of its training. ChatGPT-4o is a state-of-the-art model trained on an extensive and diverse dataset that incorporates the latest advancements in AI and healthcare knowledge, granting it enhanced interpretative and

contextual capabilities. In contrast, BioMistral, while specialized in biomedical applications, is based on relatively older architecture and limited to a narrower, domain-specific training dataset. This disparity underscores the importance of both the scale and the modernity of training data in determining the effectiveness of AI models for clinical applications. The breadth of ChatGPT-4o's training also enables it to adapt to a wider variety of scenarios, making it better equipped for real-world use in clinical decision support systems.

Clinical Implications

This research underscores the transformative potential of AI in T1D management, particularly in offering clinicians rapid, data-driven insights from CGM data. However, it also highlights that specialized training alone does not guarantee superior performance in complex, real-world applications.

ChatGPT-4o's superior baseline performance reflects its cutting-edge architecture and comprehensive training process, which includes a broad dataset encompassing diverse contexts beyond biomedical literature. This positions ChatGPT-4o as better equipped to handle complex data interpretation tasks and provide actionable insights. On the other hand, BioMistral's limitations suggest that specialization, while advantageous, must be complemented by frequent updates, larger datasets, and training processes that reflect the latest developments in both AI and healthcare. Without these improvements, specialized models risk falling behind more generalized, well-maintained systems like ChatGPT.

6.1.2 Limitations

Several limitations emerged during this study. First, the sample size was confined to pediatric patients, limiting the generalizability of findings to broader populations. Future research should encompass diverse cohorts, including adult patients with T1D, to validate the system's utility across different demographics. Additionally, BioMistral's restricted biomedical training base rendered it less effective than ChatGPT in interpreting nuanced clinical scenarios. This calls for the incorporation of larger, more diverse datasets to improve the model's comprehensiveness. Challenges related to prompt engineering and response consistency further highlighted the need for advanced optimization techniques to enhance the reliability and accuracy of specialized models like BioMistral.

6.1.3 Concluding Remarks

In summary, this thesis demonstrates that while specialized language models like BioMistral hold significant promise for healthcare applications, they require substantial enhancement to rival or surpass general-purpose models like ChatGPT. The findings advocate for a strategic approach that integrates the strengths of both specialized and general-purpose AI models to create effective, safe, and reliable clinical tools. These tools can then drive innovation in personalized, data-driven healthcare, ultimately improving patient outcomes and advancing medical practice.

6.2 Future Directions and Clinical Integration

Building on the findings of this thesis, future research should prioritize the development of specialized AI systems, such as BioMistral, to meet clinical requirements more effectively. A critical step is the enhancement of training datasets. By incorporating larger and more diverse medical datasets representing a broader spectrum of clinical cases, patient demographics, and glycemic patterns, AI models can become more comprehensive, adaptable, and capable of addressing real-world variability.

Equally important is the improvement of prompt engineering techniques. Optimizing prompt structures can enhance the accuracy, consistency, and contextual relevance of AI-generated insights, ensuring their alignment with the nuanced demands of clinical decision-making. Furthermore, the integration of real-time CGM data, alongside metrics such as physical activity and insulin dosing, offers the potential to create adaptive decision-support systems that dynamically respond to evolving patient needs.

In terms of clinical implementation, embedding advanced AI models like ChatGPT and BioMistral into Electronic Health Record (EHR) systems could enhance real-time access to patient data, improving the relevance and timeliness of generated insights. These models should also receive specialized training to recognize subtle glycemic patterns and interpatient variability, critical for diabetes care. Establishing standardized protocols for model validation and prompt engineering will be essential to ensure safe and effective use in healthcare settings.

Refining the user interface is another priority. Designing intuitive interfaces aligned with clinician workflows will encourage adoption and facilitate seamless

integration into healthcare practices. Large-scale validation studies, including multicenter trials, are also crucial to assess the effectiveness and generalizability of these systems across diverse healthcare environments.
Finally, tools like ChatGPT and BioMistral can significantly reduce cognitive load, streamline workflows, and support clinicians in making informed decisions. While ChatGPT excels in precision and detail, BioMistral's broader insights may be more suitable for less complex scenarios. To maximize their impact, these tools must operate in synergy with clinical expertise and continuously evolve to meet the demands of personalized diabetes care.

References

[1] World Health Organization. *Classification of diabetes mellitus*. Geneva: WHO, 2019.

[2] American Diabetes Association Professional Practice Committee. "2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2024". In: *Diabetes Care* 47.Supplement 1 (Dec. 2023), S20–S42. eprint: `https://diabetesjournals.org/care/article-pdf/47/Supplement\_1/S20/740359/dc24s002.pdf`.

[3] Mark A Atkinson and George S Eisenbarth. "The pathogenesis of type 1 diabetes: A scientific road map". In: *Diabetes* 50.12 (2001), pp. 2443–2470.

[4] CC Patterson et al. "Incidence trends of type 1 and type 2 diabetes among youths". In: *The Lancet Diabetes & Endocrinology* 7 (2019), pp. 517–525.

[5] American Diabetes Association. "6. Glycemic Goals and Hypoglycemia: Standards of Care in Diabetes—2024". In: *Diabetes Care* 47.Supplement 1 (2024), S111–S125. DOI: `10.2337/dc24-S006`. URL: `https://diabetesjournals.org/care/article/47/Supplement_1/S111/153951/6-Glycemic-Goals-and-Hypoglycemia-Standards-of`.

[6] CM J Nazar. "Diabetes mellitus and hypertension: A review of the interrelationship and impact on the heart and blood vessels". In: *Cureus* 9.6 (2017), e1302.

[7] Brian M Frier. "Hypoglycemia in diabetes mellitus: epidemiology and clinical implications". In: *Nature Reviews Endocrinology* 7.11 (2011), pp. 638–647.

[8] Roy W Beck, Richard M Bergenstal, and Tonya D Riddlesworth. "The challenging complexity of continuous glucose monitoring data: current and future solutions". In: *Diabetes Technology & Therapeutics* 22.S2 (2020), S1–S4.

[9] Robert S Mazze et al. "Ambulatory glucose profile: representation of verified self-monitored blood glucose data". In: *Diabetes Care* 10.1 (1987), pp. 111–117.

[10] Jan Bolinder et al. "Novel glucose-sensing technology and hypoglycemia in type 1 diabetes: a multicenter, non-masked, randomized controlled trial". In: *The Lancet* 388.10057 (2016), pp. 2254–2263.

[11] Roy W Beck et al. "Continuous glucose monitoring versus usual care in patients with type 2 diabetes receiving multiple daily insulin injections: a randomized trial". In: *Annals of Internal Medicine* 167.6 (2017), pp. 365–374.

[12] Tadej Battelino et al. "Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range". In: *Diabetes Care* 42.8 (2019), pp. 1593–1603. DOI: `10.2337/dci19-0028`.

[13] American Diabetes Association. "7. Diabetes Technology: Standards of Care in Diabetes—2024". In: *Diabetes Care* 47.Supplement 1 (2024), S126–S138. URL: `https://diabetesjournals.org/care/article/47/Supplement_1/S126/153939/7-Diabetes-Technology-Standards-of-Care-in`.

[14] J Lu et al. "Association of Time in Range, as assessed by continuous glucose monitoring, with diabetic retinopathy in type 2 diabetes". In: *Diabetes Care* 41.11 (2018), pp. 2370–2376.

[15] Louis Monnier et al. "Glucose variability: a strong independent predictor of hypoglycemia in type 2 diabetes". In: *Diabetes Care* 34.11 (2011), pp. 2317–2319.

[16] Richard M. Bergenstal et al. "Recommendations for Standardizing Glucose Reporting and Analysis to Optimize Clinical Decision Making in Diabetes: The Ambulatory Glucose Profile (AGP)". In: *Diabetes Technology & Therapeutics* 15.3 (2013), pp. 198–211. DOI: `10.1089/dia.2012.0203`.

[17] Caroline Gorst et al. "Long-term glycemic variability and risk of adverse outcomes: a systematic review and meta-analysis". In: *Diabetes Care* 38.12 (2015), pp. 2354–2369. DOI: `10.2337/dc15-1188`.

[18] Irl B. Hirsch and Michael Brownlee. "The effect of glucose variability on the risk of microvascular complications in type 1 diabetes". In: *Diabetes Technology & Therapeutics* 17.6 (2015), pp. 377–380. DOI: `10.1089/dia.2014.0327`.

[19] Richard M. Bergenstal et al. "Glucose Management Indicator (GMI): A New Term for Estimating A1C From Continuous Glucose Monitoring". In: *Diabetes Care* 41.11 (2018), pp. 2275–2280. DOI: 10.2337/dc18-1581.

[20] Jeremy Pettus et al. "The past, present, and future of continuous glucose monitoring in type 1 diabetes management". In: *Diabetes Technology & Therapeutics* 23.S3 (2021), S–3. DOI: 10.1089/dia.2021.0005.

[21] Irl B. Hirsch et al. "Role of continuous glucose monitoring in diabetes treatment". In: *American Journal of Managed Care* 25.9 Suppl (2019), S194–S203.

[22] Stephan Matthaei et al. "Consensus Recommendations for the Use of Ambulatory Glucose Profile in Clinical Practice". In: *British Journal of Diabetes* 14 (2014), pp. 153–157. DOI: 10.15277/bjdvd.2014.046. URL: https://doi.org/10.15277/bjdvd.2014.046.

[23] Robert A. Vigersky and Steven K. McMahon. "The role of continuous glucose monitoring in the management of diabetes: Evidence, insights, and recommendations". In: *Journal of Diabetes Science and Technology* 13.4 (2019), pp. 664–673. DOI: 10.1177/1932296818811573.

[24] Bruce A. Buckingham, Peter H. Chase, and Abhik Das. "Interference in Continuous Glucose Monitoring Systems: A Review of the Evidence". In: *Diabetes Technology & Therapeutics* 22.S1 (2020), S1-13-S1–19. DOI: 10.1089/dia.2019.0438.

[25] Eric Topol. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York, NY: Basic Books, 2019. ISBN: 9781541644632.

[26] Nidhi Pawar and Arjun Deshmukh. "Advances in continuous glucose monitoring for type 1 diabetes management: Review of the latest technology and applications in clinical practice". In: *Diabetes Technology & Therapeutics* 24.6 (2022), pp. 430–443.

[27] Adam Gruber, Marco Laganà, and David T. Jones. "A review of recent advances in artificial intelligence and machine learning applications for diabetes management". In: *Journal of Diabetes Science and Technology* 15.3 (2021), pp. 589–601.

[28] Shubha Mukherjee and Rahul Goswami. "Clinical decision support systems in diabetes care: A systematic review and future perspectives". In: *Journal of Diabetes Research* 2020 (2020), pp. 1–10.

[29] Sarah Ross, Helena Stephens, and Ritesh Patel. "Artificial intelligence in diabetes care: Current and future applications in continuous glucose monitoring and beyond". In: *Frontiers in Endocrinology* 13 (2022), p. 851285.

[30] Yue Liu, Jianjun Li, and Yi Wang. "Review on reinforcement learning approaches in type 1 diabetes management". In: *Computers in Biology and Medicine* 129 (2021), p. 104141.

[31] Minsu Choi, Jaewon Lee, and Katherine Smith. "Natural language processing in healthcare: Current applications and future challenges in clinical decision-making". In: *Healthcare Informatics Research* 29.2 (2023), pp. 85–98.

[32] Francesco Montani, Daniel Pereira, and Tania Chowdhury. "Evaluation of AI-driven clinical decision support systems in real-world clinical environments: A systematic review and meta-analysis". In: *BMC Medical Informatics and Decision Making* 23.1 (2023), p. 22.

[33] Abhinav Choudhury and Raunak Majumder. "Data privacy and ethical considerations in clinical decision support systems using NLP and AI". In: *BMC Medical Ethics* 24.1 (2023), p. 5.

[34] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5998–6008.

[35] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *arXiv preprint arXiv:2005.14165* (2020). URL: `https://arxiv.org/abs/2005.14165`.

[36] Pengfei Zhou et al. "Teaching Models to Generate Knowledge Statements: A Case Study in Scientific Domain". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022, pp. 738–752.

[37] C. Olsson, P. Smith, and L. Johnson. "Iterative Refinement of Language Model Prompts for Domain-Specific Applications". In: *Journal of Computational Linguistics* 49.2 (2023), pp. 123–138.

[38] Giacomo Cappon et al. "TWIN: A Digital Twin-Based Clinical Decision Support System for Type 1 Diabetes Management in Children". In: *2023 19th International Conference on Body Sensor Networks (BSN)*. Venice, Italy: IEEE, May 2023, pp. 1–4. DOI: `10.1109/BSN58485.2023.10331272`.

[39] Giacomo Cappon, Giovanni Sparacino, and Andrea Facchinetti. "AGATA: A toolbox for automated glucose data analysis". In: *Journal of Diabetes Science and Technology* (2023). DOI: 10.1177/19322968221147570.

[40] Cappon G et al. "An Integrated Mobile Platform for Automated Data Collection and Real–Time Patient Monitoring in Diabetes Clinical Trials". In: *Journal of Diabetes Science and Technology* 16.6 (2022), pp. 1555–1559.

[41] Yanis Labrak et al. "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains". In: *arXiv* 2402.10373 (2024). URL: https://arxiv.org/abs/2402.10373.

[42] OpenAI. *ChatGPT 4o*. https://openai.com. 2024.

[43] Tidepool. *Tidepool: Open-source Platform for Diabetes Data Management*. Accessed: 2024-11-13. 2024. URL: https://www.tidepool.org.

[44] G. Cappon et al. "ReplayBG: A methodology to identify a personalized model for type 1 diabetes data and simulate glucose concentrations to assess alternative therapies". Unpublished. 2024.

[45] Hugging Face. *The AI community building the future*. https://huggingface.co. Accessed: 2024-11-21. 2024.

[46] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. "Mistral 7B". In: *arXiv* 2307.08691 (2023). URL: https://arxiv.org/abs/2307.08691.

[47] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901.

[48] Jingfei Liu et al. "Can Generative Pre-trained Transformers Serve as Knowledge Bases for Closed-Book QA?" In: *arXiv preprint arXiv:2101.05208* (2021). URL: https://arxiv.org/abs/2101.05208.

[49] Maxwell C. Nye et al. "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm". In: *arXiv preprint arXiv:2102.07350* (2021). URL: https://arxiv.org/abs/2102.07350.

Acknowledgments

Concludere questo percorso rappresenta per me un traguardo di grande importanza, un momento di crescita personale e professionale che non sarebbe stato possibile senza il supporto, la guida e l'incoraggiamento di molte persone. A loro va il mio più sincero e profondo sentimento di gratitudine.

In primo luogo, desidero rivolgere un ringraziamento speciale al mio relatore, il Prof. Giacomo Cappon, per la sua preziosa guida, i suoi consigli sempre puntuali e la straordinaria disponibilità dimostrata durante ogni fase del mio lavoro. La sua esperienza e il suo entusiasmo non solo hanno orientato il mio percorso, ma sono stati per me una fonte inesauribile di ispirazione e motivazione.

Un sentito grazie va anche al Dott. Luca Cossu, correlatore di questa tesi, per il suo inestimabile supporto tecnico, la sua dedizione e il costante incoraggiamento. I suoi suggerimenti hanno arricchito profondamente questo lavoro, aiutandomi a superare le difficoltà con maggiore consapevolezza e competenza.

Un pensiero speciale va alla mia famiglia, in particolare ai miei genitori che dedico queste parole: আমার প্রিয় বাবা-মা,
এই সাফল্যের পিছনে তোমাদের অবদান এবং নিরন্তর ভালোবাসার জন্য আমি চিরকাল কৃতজ্ঞ। তোমাদের সমর্থন, ত্যাগ এবং অনুপ্রেরণা ছাড়া আজকের দিনে পৌঁছানো কখনোই সম্ভব হতো না। তোমরা সব সময় আমার পাশে থেকেছো, আমাকে সাহস জুগিয়েছো, এবং আমার স্বপ্নকে তোমাদের স্বপ্ন বানিয়েছো। আমার জীবনের প্রতিটি পদক্ষেপে তোমাদের ভালোবাসা, উৎসাহ এবং বিশ্বাস আমার শক্তি হয়ে কাজ করেছে।
তোমরা আমার জীবনের সবচেয়ে বড় দোয়া। আমার হৃদয়ের গভীর থেকে তোমাদের প্রতি জানাই অশেষ ভালোবাসা এবং কৃতজ্ঞতা।.

Un grazie va a mio fratello e alla mia bhabhi che mi hanno sostenuto con amore e pazienza lungo tutto il percorso universitario. Senza la loro vicinanza e il loro sostegno incondizionato, raggiungere questo obiettivo non sarebbe stato possibile.

Grazie di cuore ai miei amici, che hanno saputo rendere più leggero e gioioso il cammino con momenti di condivisione, conforto e allegria. Un grazie particolare agli amici di sempre, che mi hanno accompagnato fin dall'inizio.

Un ringraziamento particolare va a una persona speciale che, con il suo affetto, la sua comprensione e la sua costante presenza, ha reso questo percorso più sereno. La sua capacità di spronarmi nei momenti difficili e di gioire con me per ogni piccola vittoria ha fatto una differenza fondamentale.

Un riconoscimento speciale va agli amici della Biblioteca di Silea, che sono stati al mio fianco nei momenti più impegnativi, offrendomi supporto e incoraggiamento. Grazie anche ai compagni universitari che ho incontrato lungo il percorso: a chi mi ha accompagnato fin dai primi anni e a chi ho conosciuto negli ultimi momenti nel meraviglioso regno della SE o nel suo modesto sostituto, la GE. Un ricordo particolare va alle pause condivise in area lounge, piccoli attimi di leggerezza che resteranno impressi nella memoria.

Un pensiero speciale va poi agli amici conosciuti durante i viaggi in treno verso Padova, che hanno trasformato quelle lunghe ore di spostamenti quotidiani in occasioni di dialogo, condivisione e amicizia. Con voi il viaggio è stato non solo un percorso fisico, ma anche umano, arricchito da risate, riflessioni e momenti di complicità.

Non posso dimenticare il mio "sponsor" non ufficiale: Piovego, che con i suoi deliziosi pasti mi ha aiutato a superare le giornate più difficili nella frenetica vita padovana.

Grazie a tutti coloro che, in modi grandi e piccoli, hanno contribuito a rendere possibile la realizzazione di questo lavoro.

Infine, un ringraziamento a me stesso, per aver creduto in questo obiettivo e per la determinazione che mi ha permesso di superare le difficoltà e raggiungere questo importante traguardo.