Computer Methods and Programs in Biomedicine 269 (2025) 108878

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

A novel fine-tuning and evaluation methodology for large language models on IoT raw data summaries (LLM-RawDMeth): A joint perspective in diabetes care

Juan F. Gaitán-Guerrero [a,*], Carmen Martínez-Cruz [b], Macarena Espinilla [a], David Díaz-Jiménez [a], Jose L. López [a]

[a] Department of Computer Science, University of Jaén, Jaén, 23071, Spain
[b] Department of Languages and Computer Systems, University of Granada, E.T.S. de Ingenierías Informática y de Telecomunicación, Granada, 18071, Spain

ARTICLE INFO

Keywords:
IoT-data fuzzy summarization
Large language models
Fine-tuning process
Continuous glucose monitoring
Prompt engineering
Evaluation methodology

ABSTRACT

Background and objective: Diabetes is a global health concern, affecting millions of adults worldwide and exhibiting a growing prevalence. Managing the disease highly relies on continuous glucose monitoring, yet the dense and complex nature of electronic devices data streams poses significant challenges for efficient interpretation. Large Language Models are being widely applied across different domains for their ability to generate human-like text, but still fall short in producing accurate and meaningful text from raw data. To address this limitation, this study proposes a fine-tuning methodology tailored specifically to glucose data, but scalable to other expert-guided domains, enabling the models to generate concise, relevant and safe summaries, bridging the gap between raw data and efficient medical attention.

Methods: This study introduces a novel continuous glucose monitoring framework that involves fine-tuned GPT models using structured datasets generated through an expert-guided data modeling based on Fuzzy Logic and prompt engineering for task contextualization. A new evaluation methodology is defined to assess the performance of the Large Language Models across different critical domains where expert knowledge is fundamental to characterize temporally dependent data and ensure valuable insights.

Results: Fine-tuned GPT-4o achieved the highest performance, with an average score of 96% across all metrics. GPT-4o-mini followed with 76% score, while GPT-3.5 scored 72%. The use of fuzzy knowledge-based prompts proved more effective in scenarios with full data availability, or in scenarios with a simplified data availability when the models are not fine-tuned; domain-guided prompts improved output relevance and stability in fine-tuned models with less data availability.

Conclusions: These results indicate the capability of our methods to align Large Language Models with the task of generating human-like text from raw data, highlighting their potential to manage diabetes by complex glucose patterns interpretation, alleviating the burden on healthcare systems.

1. Introduction

Since the mid-20th century with Alan Turing's test, human beings have the need to develop computational intelligent behaviors and evaluate them to see how similar and indistinguishable they are from us. Nowadays, the applications of Artificial Intelligence (AI) have reached all areas of knowledge, with the aim of assisting humans in their needs [1]. One of the applications that has gained the most attraction in recent years is the use of Language Models, particularly Large Language Models (LLMs), which are very close to being an Artificial General Intelligence that can be applied to any field [2].

LLMs are focused on the processing and interpretation of text to generate content in natural language. These AI models [1] are being used to solve general purpose linguistic tasks and the research paradigm has shifted towards its use. It is important to note that the scientific community has developed several pre-trained LLMs [1], being both GPT-3.5 and GPT-4o (developed by OpenAI) particularly important in the current literature [3–10] due to their promising results and fast response, together with the no need for dedicated and proprietary technological infrastructure. In contrast to most open-source models, OpenAI provides an integrated working interface that simplifies access

* Corresponding author.
E-mail addresses: jgaitan@ujaen.es (J.F. Gaitán-Guerrero), cmcruz@ugr.es (C. Martínez-Cruz), mestevez@ujaen.es (M. Espinilla), ddjimene@ujaen.es (D. Díaz-Jiménez), llopez@ujaen.es (J.L. López).

https://doi.org/10.1016/j.cmpb.2025.108878
Received 25 January 2025; Received in revised form 12 April 2025; Accepted 26 May 2025
Available online 9 June 2025

and usage, combining computational power, usability and a broader user reach. Up to this point, it is important to note that OpenAI has top 400 million users despite the publication of other latest publicly available models [11].

1.1. General and healthcare applications of LLMs

As a consequence, LLMs are being applied in a wide variety of domains [1], such as education [12], legal analysis [13,14], agriculture [15], or in scientific research, where LLMs demonstrate effectiveness in handling knowledge-intensive tasks and in developing field-specific models [9].

In the healthcare context, LLMs are a powerful tool for both patients and professionals, as attested by the general approach presented by Campillos-Llanos et al. [6], which proposes LLMs to support clinical text mining. In the same line, Sciannameo et al. [10] propose InstructGPT to bypass task-specific training in information retrieval from unstructured natural language text. Furthermore, on [4] the authors proposed a methodology for mental health analysis utilizing LLMs. On the other hand, Nov et al. [5] demonstrate that LLM-generated medical advice can be indistinguishable from healthcare professionals responses.

1.2. Integrating LLMs with IoT

Building on these advancements, the integration of LLMs with IoT data is found pivotal in opening new opportunities for healthcare applications. IoT data, which comes in the form of time series (TS), is dense and difficult for LLMs to interpret [16,17]. It is therefore a challenge to align TS and natural language in order to take advantage of their capabilities. Approaches like prompt-as-prefix (PaP) methodology, consist in enriching the input TS with additional context and providing instructions for the model in classification, regression or reasoning tasks on IoT-streamed data. The HarGPT study [18] and the LLaSA model [19] shows the potential of LLMs to interpret and classify sensor data for human activity recognition. A more ambitious approach is presented by An et al. [20] to leverage LLMs for IoT task reasoning and classification, enhancing their capabilities through tailored data preprocessing, IoT-specific knowledge augmentation and optimized prompt design. This work again mentions LLMs to be struggling with complex TS data, specially when coming from physiological signs, suggesting fine-tuning (FT) to these modalities due to data intricate temporal patterns and the domain-specific language required for accurate interpretation. In this sense, a large number of works [4,6,8–10,21,22] have shown that models can improve their capabilities by adapting them to specific tasks through the process of FT. Nonetheless, in the literature we can find a large number of papers that apply this technique in very different ways, without a scientific consensus [1,21]. For instance, in the IoT domain, Liu et al. [23] propose instruction tuning with domain-specific prompts for accurate blood pressure estimation from wearable devices data.

1.3. Linguistic descriptions of IoT data in clinical monitoring

However few approaches address the generation of linguistic descriptions of IoT-data on behalf of LLMs. In fact, in clinical contexts requiring continuous monitoring, such as chronic diseases, there is a growing need to translate the complex numerical data from IoT devices into interpretable linguistic summaries. Recent approaches like Penetrative AI [24] and the work of Fang et al. [25] pretend to address this topic by preprocessing raw signals into simplified, textual summaries, or even utilizing context-driven prompts and statistical analysis, without the models directly handling unstructured numerical data. This methods enhance interpretability but remains limited in scalability and real-time raw data processing.

Among chronic diseases requiring continuous monitoring, diabetes stands out as a critical case due to the need for personalized treatment adjustments, a challenge further amplified by its global prevalence (with an estimated 643 million adults affected by 2030) [26] and the growing burden it places on healthcare systems worldwide. Nonetheless, new devices for continuous monitoring of glucose from the interstitial fluid, which is generated from exchanges between tissue cells and blood, have appeared. These devices, such as the Dexcom [27] or the Freestyle Libre [28] sensors, establish communication with mobile devices, allowing patients to improve their quality of life by having uninterrupted control of their glycemia dynamics. The Freestyle Libre sensor is the most widely used in its third version [29], which is characterized by its long battery life (14 days), a continuous sending of glucose data (every 5 s), its small size (21×2.9 mm) and the ease with which it can be attached to the body. As a consequence, a huge flow of information is generated over time for each patient. Combined with the fact that most medical appointments are time-restricted, the evaluation of all required data is limited for a truly personalized care, since there is no centralized framework where patients can provide their data to obtain an analysis. Therefore, there is a clear need for support systems that collect all the measures captured by these devices, and presents a summary in human-like language that is adequate, fast and safe, highlighting the most relevant events that have occurred, for both healthcare professionals and patients. In the absence of such centralized framework, it is highlighted that some patients may utilize publicly available LLMs to interpret their own data, even though these models are not yet clearly able to directly handle raw TS data.

1.4. LLMs in the domain of diabetes

In the context of diabetes and leveraging LLMs, M. Abbasian et al. [22] propose an agent to calculate food intake in relation to the guidelines of the American Diabetes Association. This agent is designed with electronic health records of diabetic patients to provide nutritional guidelines. On the other hand, the work of D. Dao [30] proposes a methodology based on LLMs for the prevention of diabetes, reducing the risk factors of this disease. To achieve this, activity data is collected from the user's smartphone and calendar to provide alerts, reminders and personalized suggestions to encourage a healthy lifestyle. However, although dietary habits and daily activity are essential to maintaining good health in people with diabetes, these systems do not include active monitoring of patients with diabetes to support healthcare professionals in their daily tasks, for example by monitoring blood glucose levels using linguistic summaries. Building on this need, a preliminary study is presented in [31], which analyzes the results of GPT-4 in glucose monitoring data summarization, addressing a study of the required input information and the expected output template, thus highlighting the need to develop adapted LLMs to improve the results, as the given 14-day summarization is not fully reliable, thereby necessitating a shift towards a daily summarization approach as a foundational step before scaling to longer periods. Besides, only the medical staff is contemplated in the evaluation, without considering all the potential users of the system, being the score conditioned by the expertise and knowledge of the healthcare personnel. In the same line, Healey et al. [32] proposes different questions to be answered objectively on behalf of LLMs about continuous glucose monitoring (CGM) aiming to establish a benchmark to evaluate them.

1.5. Output evaluation of LLMs

Focusing on evaluation procedures, traditional metrics like BLEU, ROUGE, or BERTScore are often used in summarization tasks, but they fall short when applied to time series (TS) analysis, where capturing relevant details and reasoning is essential. Recent studies emphasize the need for more holistic evaluation frameworks that combine computational and expert perspectives [33,34], especially in healthcare, where accuracy and safety are critical. While works like [25,31] rely on survey-based human evaluation – either without a defined ideal output

or by embedding the template in the prompt – Yang et al. [4] highlight the importance of a structured, domain-specific framework. Given the complexity of TS data, where events may repeat or vary in form and timing, existing semantic or event-based metrics are limited. As a result, there is still no scientific consensus on how to reliably evaluate LLMs in TS description tasks, and survey-based methods remain subjective and incomplete.

According to the latter, the taxonomy proposed by Fons et al. [35] addresses the evaluation of TS descriptions performed by LLMs, by detecting key features, identifying critical points, and assessing arithmetic reasoning. However, it lacks complementary metrics to interpret the models' behavior, viability and robustness.

To a broader extent, the QUEST framework [36] offers a general structure for human evaluation in healthcare, focusing on principles like information quality, reasoning, expression style, and trust. However, it does not consider the challenges of evaluating LLMs with raw data streams, such as those from IoT-based patient monitoring. Since it is centered on the output and the evaluator's opinion, it often uses Likert scales, which makes the process subjective and dependent on trained annotators. This limits the possibility of reproducible and scalable evaluations. The framework also ignores the importance of outputs that follow a clear and consistent temporal evolution—something essential not only in TS descriptions but also in clinical contexts. Even when adapted to specific tasks, the way the metrics are defined does not allow for a clear calculation of whether the model is truly providing the expected information, beyond what seems correct to the evaluator.

1.6. Advancing limitations in the literature

To synthesize the current state of the art and identify the limitations found in the literature, a summarization of the reviewed works is presented in Table 1. This horizontal comparison aims to highlight the key dimensions across which the existing works differ, revealing that only a few proposals enable LLMs to process raw TS data directly, without prior preprocessing, thus restricting the applicability of such models to real-world streaming data scenarios. Secondly, although FT is frequently mentioned as a promising direction, very few works implement it, and none apply it for data description. Furthermore, there is no consensus on the evaluation of LLMs through a generic and scalable framework that could be applied either automatically or by non-domain experts, highlighting that even general-purpose frameworks do not consider metrics associated with temporally dependent data scenarios. Finally, Table 1 reflects the predominance of GPT-series models in the reviewed works.

As a result, this work introduces the following key innovations:

1. Creation of a framework for expert knowledge modeling of raw data, enabling the obtaining of a training dataset for adapting LLMs through a FT process.
2. Exploration of prompt engineering strategies, showing how tailored prompts improve the output of LLMs in specialized tasks.
3. Definition of a standard methodology for robust evaluation of LLMs when handling raw temporally data (LLM-RawDMeth), with the inclusion of metrics to be obtained objectively.
4. Demonstration of the capability of GPT technology to adapt to specific domains through a FT process on limited datasets.

To the best of our knowledge, this is the first system of its kind to meet natural language generation standards for raw IoT data by proposing a dual innovation: a methodology that jointly addresses raw data modeling (expert committees and international standards) and the definition of a evaluation framework that is both objective and accessible to non-experts, without relying on extensive human supervision.

The proposed process is organized as follows: Section 2 presents a new methodology for adapting and evaluating LLMs' performance in the specific domain of diabetes. The obtained results from the case study in utilizing GPT models are exposed in Section 3 and discussed in Section 4. Finally, Section 5 concludes the study.

2. Methods

This section narrates the methodology of the system (Section 2.1) which exposes the procedure followed for the development of a centralized monitoring framework that allows for data acquisition and communication with LLMs for its summarization. Furthermore, this section also contemplates the design of an evaluation framework to determine LLMs performance in Section 2.2.

2.1. From IoT to LLMs: Methodology of the system

This section is divided into three different phases to better understand the different stages that compose it. In this way, our methodology first contemplates the extraction of patients' glucose levels via an IoT-driven architecture in Phase 1 (Section 2.1.1), which involves the interconnection of the different devices, servers and their components. Once data is collected, Phase 2 (Section 2.1.2) is applied for the generation of a dataset, by the definition of relevant observable phenomena to model the collected values, on behalf of medical staff criteria. This process holds data preprocessing, data labeling, segmentation for pattern identification and, the generation of a final output considering predefined templates and quality rules. Subsequently, Phase 3 (Section 2.1.3) corresponds to the creation of a structured JSONL file whose messages represent a conversational scenario for describing each TS; a latter dataset splitting and hyperparameters' setting experimentation allow for the creation of a specialized model in the domain being handled. Additionally to these three stages, Phase 4 (Section 2.2) is presented to enable the evaluation of linguistic summaries, which are generated by the specialized model varying the prompt provided; both the prompt designing and the evaluation metrics are designed by experts to assess the performance of the system. The entire process is illustrated in Fig. 1. To test the validity of this methodology we test its use in a real environment, in the context of diabetes disease monitoring.

2.1.1. Phase 1. Architecture of the system

Diabetic patients have recently taken advantage from new commercial devices that continuously monitor glucose levels. These are considered wearable devices due to their subcutaneously placement. In this work, the Freestyle Libre 3 sensor [29] is used. Data is retrieved from the sensor through the utilization of Bluetooth Low Energy (BLE) connection and Near-Field Communication (NFC) pairing, with a smartphone device running a mobile application to facilitate this task, xDrip+ [37] in this case, facilitating real-time monitoring tasks.

The data is then sent to a proprietary server for storage and processing. A RESTful API is configured to manage this process and to retrieve data from a MongoDB-powered database in order to prepare it to serve as input for the LLM. Note that a 5-min interval separates each glucose sample, deriving in a total of around 288 samples a day. Table 2 show the attributes that characterize every instance of a patient's dataset. It is important to note that only the timestamp, the utcOffset and the sgv attributes are considered. The timestamp denotes the date and time the sample was registered, which is adjusted according to the utcOffset; the sgv attribute represents the glucose value obtained. As observed, no personal data that could identify the patient is stored in the database, nor is it provided to the LLM. The entire dataset is available in [38].

This IoT architecture has allowed for the obtaining of a complete overview of patients' glycemia. Therefore, considering the deep cognitive effort the healthcare personnel must put in analyzing every patient fluctuation, an expert committee has helped to model the most interesting and needed information to be obtained from the TS.

Table 1
Horizontal comparison of the reviewed approaches, focused on the healthcare realm, in the literature

Proposal	Domain	Input data type	Model series	Model adaptation	Main goal	Type of evaluation
Campillos-Llanos et al. [6]	Clinical text mining	Text	RoBERTa	Fine Tuning	To obtain structured information from unstructured medical text	Precision, Recall, F1 score and human evaluators
Sciannameo et al. [10]	Clinical text mining	Text	GPT	Instruction Tuning	Information retrieval	Precision and human annotators
Hu et al. [4]	Clinical text mining	Text	Qwen	Fine Tuning	Case analysis in mental health	Accuracy, ROUGE, BLEU and BERTScore
Nov et al. [5]	Q&A	Text	GPT	Prompt-guided	Medical advice generation	Human evaluators
Ji et al. [18]	Human activity recognition	Raw data	GPT	Prompt template	Prediction of activities based on sensor data	Precision, Recall and F1 score
Imran et al. [19]	Q&A and human activity recognition	Raw data (embeddings)	LLaMA	Fine tuning	Prediction of activities based on sensor data	Precision, Recall, F1 score, Correctness, Completeness, Consistency, Helpfulness and human evaluators
An et al. [20]	Human activity recognition	Enriched raw data	GPT, Claude, Gemini, Mistral, LLaMA	Prompt-guided	Prediction of activities based on sensor data and description of the decision	Accuracy, MAE, STD and human evaluators
Liu et al. [23]	Physiological data	Enriched raw data	Qwen, Gemini, Mistral, LLaMA	Instruction tuning	Prediction of blood pressure	MAE, ME and SDE
Xu et al. [24]	Physical world data	Enriched raw data	GPT	Prompt-guided	Prediction and description in comprehending physical world through sensor data	Precision, MAE and comparison with other methods
Fang et al. [25]	Physiological data	Enriched raw data	GPT	Instruction tuning	Health insights based on physiological data	Likert scale and interviews
Abbasian et al. [22]	Diabetes management	Text	GPT	Prompt-guided	Nutritional intake calculation	Precision and comparison with other methods
Dao et al. [30]	Diabetes management	Text	GPT	Fine tuning	Diabetes prevention	Human evaluators
Healey et al. [31]	Diabetes management	Raw data	GPT	Prompt-guided	Description of CGM data	Ground truth, accuracy, completeness, safety and usability
Healey et al. [32]	Diabetes management	Raw data, text, code	GPT	Prompt-guided	Description of CGM data through Q&A	Ground truth and human evaluators
Fons et al. [32]	TS analysis	Raw data, text	GPT, LLaMA, Vicuna, Phi	Prompt-guided	Analyze the capability of LLM to interpret and analyze TS	F1 score, accuracy and MAPE
Tam et al. [36]	Evaluation of LLM	–	–	–	To define a framework for human evaluators	Multiple (subjective)
Our approach	TS analysis and Evaluation of LLM	Raw data	GPT	Prompt-guided and Fine tuning	Analyze the capability of LLM to interpret and analyze TS	Multiple (objective), adding the metric of temporal dynamics

2.1.2. Phase 2: Dataset generation

The intrinsic need to understand what is expected to be taken as output from the input data derives in expert knowledge to take a crucial role in defining the most relevant characteristics, events or episodes to be highlighted. These are the aspects experts take a deep look at, when suggesting treatment changes and/or recommendations for a patient whose glycemia is being controlled for their well-being, aiming to better manage the disease.

In this regard, the events that are particularly prevalent in the domain of diabetes care are when peaks or valleys (the latter ones commonly referred to as hypoglycemia episodes) are detected. In parallel, a distinction can be made between a peak or a hyperglycemia episode, the latter being characterized by a rise in glucose levels and

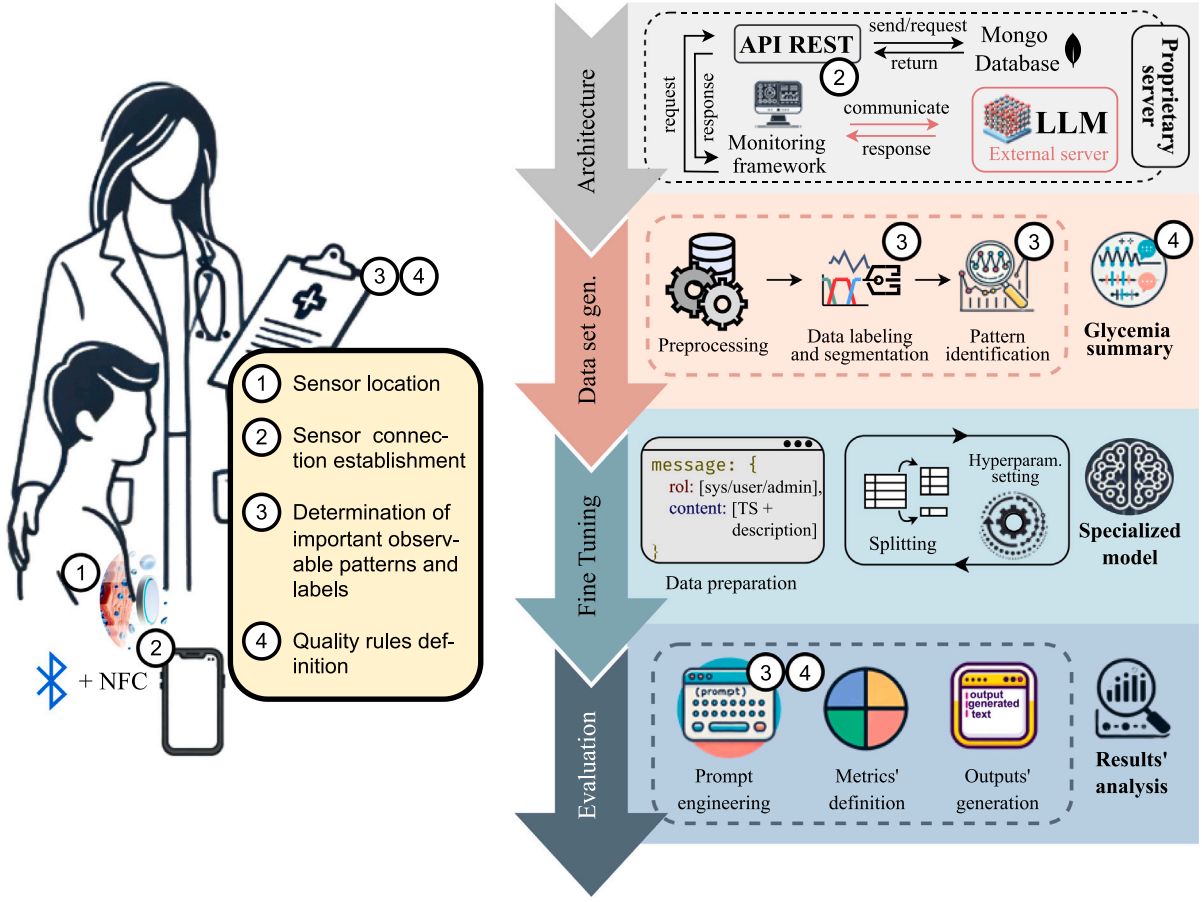

Fig. 1. Methodology to adapt IoT data to LLMs systems.

Table 2
Sensor data instance example.

Field	Value
date	1711408157452
dateString	2024-03-25T23:09:17.434Z
sgv	84
delta	8.007
direction	FortyFiveUp
type	sgv
filtered	127 000
unfiltered	127 000
rssi	100
noise	1
sysTime	2024-03-25T23:09:17.434Z
utcOffset	60

Table 3
Labels membership functions for glucose values.

Label	Trapezoidal function
Very low	*z-shape*$(-\infty, -\infty, 54, 60)$
Low	*trapmf* $(54, 60, 75, 80)$
Medium	*trapmf* $(75, 80, 120, 130)$
High	*trapmf* $(120, 130, 160, 170)$
Very high	*s-shape*$(160, 170, \infty, \infty)$

Table 4
Labels membership functions for segment trend.

Label	Trapezoidal function
Sharply decreasing	*z-shape*$(-\infty, -\infty, -0.75, -0.5)$
Decreasing	*trapmf* $(-0.75, -0.5, -0.25, -0.1)$
Steady	*trapmf* $(-0.25, -0.1, 0.1, 0.25)$
Increasing	*trapmf* $(0.1, 0.25, 0.5, 0.75)$
Sharply increasing	*s-shape*$(0.5, 0.75, \infty, \infty)$

their prolongation over time. Likewise, it is important to denote those time intervals characterized by glucose values that are within a given threshold, enabling possible pattern establishment in the behavior of a patient's glucose levels to provide more personalized care (e.g., indicate that glucose values appear to be under 70 mg/dL in the morning).

For building the described knowledge, it is found as imperative to take advantage of Fuzzy Logic theory proposed in [39], which is incorporated within the Knowledge Discovery in Databases procedure for the utilization of a natural language generator, as detailed in Marín & Sánchez [40]. These paradigms allow for the representation of implicit information that characterize a knowledge which remains unknown until posterior consideration. Essentially, the variables modeled in a fuzzy way in this proposal are those related to moments of the day (L_R, Table 5), glucose values (L_S, Table 3), trend between segments (L_T, Table 4), and quantifiers (Q, see Table 6), that allow associating the occurrence of certain glucose levels in a particular time interval of the day. To do this, it is necessary to use type 2 prototypical forms (protoforms), which are structured in the form R (day moment) Q (quantifier) *glucose values are* S (glucose values summarizer). The relevance of the generated protoforms is determined by the Degree of Truth (DoT), utilizing the expression given in Algorithm 1.

In order to obtain the maximum number of points to be analyzed from the TS, both for the calculation of protoforms and for obtaining segments for the identification of patterns, the Moving Average Window algorithm is used to estimate possible data loss over an interval of up to half an hour. The result is the creation of protoforms and the consideration of different sets of consecutive segments for the

Table 5
Labels membership functions for day intervals.

Label	Trapezoidal function
Night	*z-shape*$(-\infty, -\infty, 6\text{ am}, 8\text{ am})$
Morning	*trapmf*$(6\text{ am}, 8\text{ am}, 12\text{ pm}, 2\text{ pm})$
Afternoon	*trapmf*$(12\text{ pm}, 2\text{ pm}, 8\text{ pm}, 10\text{ pm})$
End of day	*s-shape*$(8\text{ pm}, 10\text{ pm}, \infty, \infty)$
During the daytime	*s-shape*$(7\text{ am}, 9\text{ am}, \infty, \infty)$

Table 6
Specification of quantifiers.

Quantifier	Trapezoidal function
Few	*s-shape*$(10, 30, \infty, \infty)$
Many	*s-shape*$(40, 60, \infty, \infty)$
Most	*s-shape*$(60, 80, \infty, \infty)$
Almost all	*s-shape*$(80, 100, \infty, \infty)$

identification of possible pre-defined patterns (SPD). It is important to emphasize that the final summary is subject to a series of quality standards with the aim of obtaining a clear, concise and non-redundant output, so that the most relevant information is summarized. Therefore, the following objections are taken into account:

- TS descriptions progress from general description to specific phenomena.
- The summary should be as concise as possible.
- Actual time measurements are rounded to the nearest quarter-hour, and time periods are approximated to the closest hour.
- Descriptions of very short periods (less than one hour) are omitted.
- Consecutive segments or sentences with similar descriptions are combined.

The whole procedure described in this section is provided in more detailed in Algorithm 1. Note that each point x_t in the TS is characterized by the sensor glucose value ($x_{t_{sgv}}$), the associated linguistic label ($x_{t_{SLabel}}$) and the corresponding membership degree ($x_{t_{SValue}}$) which is computed using the trapezoidal membership function μ. Idem for the characterization of the moment of the day to which the point belongs ($x_{t_{RLabel}}$, $x_{t_{RLabel}}$). The trend of a segment in the segmented TS s_{TLabel} is also calculated through the membership function μ, being the value to compute (s_{TValue}) the trend between the start and the end point of the segment. Additionally, it is highlighted that $x^{\{i,j\}}$ correspond to a set of consecutive segments.

An example for the generated linguistic description for graph (d) in Fig. 3 is shown next: *The collected values comprehends the whole day and no data was lost. The patient experienced many low values during the whole day. At night, a hypoglycemia episode of 39 mg/dL was detected at around 3:00 am and lasted for 2 h. In the morning, the patient experienced few low glucose values and a hypoglycemia episode of 38 mg/dL was detected at around 10:00 am. In the afternoon, a peak of 142 mg/dL was detected at around 3:45 pm, followed by a hypoglycemia episode of 53 mg/dL at around 5:00 pm. At the end of the day, a hypoglycemia episode of 56 mg/dL was detected at around 8:45 pm. At last, the patient experienced few low and few very low glucose values with an increasing trend.*

2.1.3. Phase 3: Fine tuning procedure

The computation of different TS allow the creation of a dataset that serves as the training data for a specific LLM, leading the specific model to adapt itself for the specific desired task. To do the FT task, the dataset is formatted as a JSONL file using OpenAI's structured roles system i.e.: the *system* role for providing initial instructions and context for the task, the *user* specifying the TS data, and the *assistant* role with the expected output. The format is specified in [41] and shown next:

Algorithm 1 Glucose Data Characterization and Linguistic Summarization

Input: Glucose TS (**TS**)
Input: Sets of labels ($\mathbf{L_R}, \mathbf{L_S}, \mathbf{L_T}$)
Input: Quantifiers **Q**
Input: Segment pattern description (**SPD**)
Output: Linguistic summarization of data (**LS**)

Gap filling and point labeling
for $t \leftarrow 1$ to $|TS|$ do
  if $x_{t_{sgv}} = \emptyset$ and $|x_{t:t+6}| < 6$ then
    $x_{t_{sgv}} \leftarrow \text{MovingAverageWindow}(x_{t-3}, x_{t+3})$
  end if
  if $x_{t_{sgv}} \neq \emptyset$ then
    $x_{t_{SValue}} \leftarrow \max(\mu(x_{t_{sgv}}, [s_a, s_b, s_c, s_d]), \forall s \in L_S)$
    $x_{t_{SLabel}} \leftarrow s : \max(\mu(x_{t_{sgv}}, [s_a, s_b, s_c, s_d]), \forall s \in L_S)$
    $x_{t_{RValue}} \leftarrow \max(\mu(x_{t_{tstamp}}, [r_a, r_b, r_c, r_d]), \forall r \in L_R)$
    $x_{t_{RLabel}} \leftarrow r : \max(\mu(x_{t_{tstamp}}, [r_a, r_b, r_c, r_d]), \forall r \in L_R)$
  end if
end for
Generation of protoforms
$P \leftarrow \emptyset$
for $x^r \subset TS\ \forall r \in R$ do
  for $x^s \subset TS\ \forall s \in S$ do
    $q' \leftarrow 0$
    for $q \in Q$ do
      $DoT \leftarrow f(\text{sum}(x^r_{i_{RValue}} \wedge x^s_{i_{SValue}})/\text{sum}(x^r_{j_{RValue}}), \forall i \in 1\ to\ |x^r \cap x^s|, \forall j \in 1\ to\ |x^r|, [q_a, q_b, q_c, q_d])$
      if not $DoT < 0.7$ then
        $q' \leftarrow q$
      end if
  end for
  $P \leftarrow P \cup \{(r, q, s)\}$
end for
Time Series Segmentation
$\epsilon \leftarrow 0.2$
$segmented_{TS} \leftarrow \text{RamerDouglasPeucker}(TS, \epsilon)$
for $s \leftarrow 1$ to $|segmented_{TS}|$ do
  $s_{TLabel} \leftarrow \max(f(s_{TLabel}, [t_a, t_b, t_c, t_d]), \forall t \in L_T)$
end for
Pattern identification
$W \leftarrow \emptyset$
for $x^{\{i,j\}} \subset segmented_{TS}$ where $i < j$ and $i \cap j \neq \emptyset$, $\forall i, j \in 1\ to\ |segmented_{TS}|$ do
  if $x^{\{i,j\}} \subset SPD$ then
    $W \leftarrow W \cup x^{\{i,j\}}$
  end if
end for

$LS \leftarrow e$ for $e \in P \cup W$ if $qualitySatisfaction(e)$

```
{
 "messages": [
     {
     "role": "system",
     "content": " [Domain], [Instruc-
       tion] and [Knowledge]"
   },
   {
     "role": "user",
     "content": "[Petition] <TS>
       (format[HH]:[MM]am/pm -
       [sgv]mg/dL)"
```

```
    },
    {
      "role": "assistant",
      "content": "[Expected output
      summary of glucose TS]"
    }
  ]
}
```

Phase 4 consists in the evaluation of the designed system. For this purpose, we first define the evaluation mechanism (Section 2.2) on which the experiments will be performed (Section 3).

2.2. A new LLMs evaluation methodology (LLM-RawDMeth)

In this section, the proposed methodology for evaluating the behavior of LLMs when summarizing raw data is presented. The profound concerns emerged after reviewing the existing literature has derived in the design of a deep study that better contributes in domains with temporally dependent data (numerical or textual), being the domain of diabetes our case study.

The result of the previous expert-guided linguistic summarization emerges as a powerful tool for analyzing the collected glucose data. Taken together, the expert committee proposal described in Section 2.1.2 and the international consensus group guidelines described in [42] for the interpretation of diabetic-patient fluctuations, are found as the premises of this in-depth study.

Firstly, the prompt design, is contemplated due to the relevance of this technique in the field of natural language processing. For this reason, the proposed methodology focuses on the design and optimization of the prompt used when interacting with LLMs; the recent advances in LLMs have evidenced the remarkable superiority of prompt engineering in several areas [43], highlighting the need to identify the key elements it should include. In fact, OpenAI [44] also specifies the main strategies an user should follow when typing clear instructions for the model. The steps to follow involve defining the level of detail in the query, instructing the model to adopt a persona, using delimiters to separate distinct parts of the input, and determining the desired length of the response. Other common tactic is to provide examples of the expected output and/or its format, which has been discussed in [16], where the idea of a PaP is given to avoid this task.

Taking all aspects into consideration, our prompt design consists in 4 different parts. The first one introduces the input information letting the LLM to understand the task to make. Next, the TS is provided utilizing a delimiter to separate it from the rest of the prompt. Secondly, the domain of the data is described emphasizing in the aspects to be obtained from the raw data. To do so, in the third part, a set of instructions is defined to ask the model to adopt a determined role, produce a specific output length ($\langle L \rangle$) and establish the time interval ($\langle T \rangle$) to be analyzed. In the knowledge part, the prompt focuses on establishing the different labels and metrics the LLM must utilize for data computation ($\langle G \rangle$ for glucose levels, $\langle D \rangle$, for day intervals, and $\langle M \rangle$ for quantifiers). The entire prompt can be observed in Fig. 2. Note that no pre-defined template is considered as the exposed process deals with the utilization of training data to adapt the used model.

Consequently, different TS are being evaluated according to different criteria basis. The assessment of the performance of different LLMs in satisfying the anticipated human expectations is carried out by scoring each output in terms of the criteria and the associated metrics defined for them, as specified in Table 7. The evaluation framework outlined provides a detailed and objective methodology for assessing the performance of a LLM in the expert-guided and/or critical domains with temporally dependent data across four key dimensions (based on [36]): Information Quality, Thought Capabilities, Communication Quality, and Content Safety. First, Information Quality pretends to evaluate every statement that is present in the output generated, with the objective of determining if they meet the ideal message. Next, the Thought Capabilities are evaluated across one single metric that consider the understanding of the input on behalf of the LLM, together with the application of logic and thinking when generating the response. According to the latter, Communication Quality metrics are designed aiming to ensure the language is appropriate and understandable, while also accounting for the temporal characteristics of the data. Finally, Content Safety criteria ensure the accuracy of the linguistic output while avoiding unnecessary or unsupported statements derived from the data, which could potentially influence end-user behavior negatively.

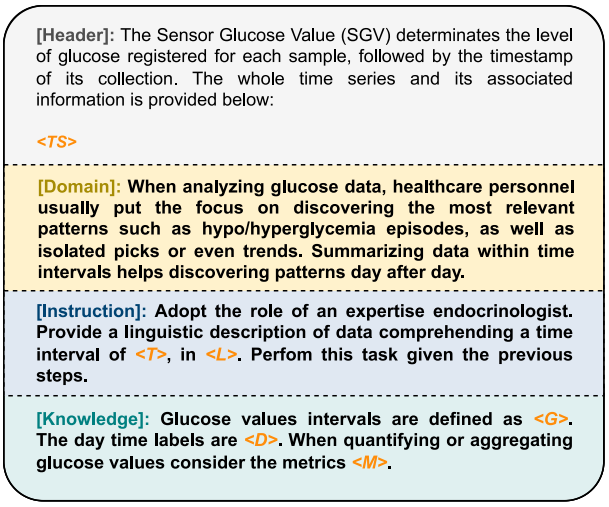

Fig. 2. Prompt structure design based on [16] and a diabetes management expert-guided procedure. The schema elements denoted as <> allow individual personalization.

Under Information Quality, *Acc*, *Re*, *Ct* and *R* ensure the non-appearance of incorrect, superfluous or unexpected knowledge; *Rp* measures the stability of responses over time. Thought Capabilities are evaluated by assessing the model's comprehension of the query and the logical coherence of its responses (*Und*). Communication Quality evaluate *C*, *Ab* and *Td* for conciseness, comprehensible language, and accurate time referencing. Lastly, Content Safety involves *H* and *Av* checking for the absence of false statements, the model's recognition of its limitations and the procedures used, and for ensuring responses are grounded in the input data without unfounded facts.

3. Results

In this section, the proposed methodology has underscored the validity of our proposal, positioning expert knowledge modeling as an essential need to empower the emerging technologies in the well-known as AI era. To do so, a set of 4 different and varied TS has been considered, which are graphically presented in Fig. 3.

3.1. Experimental data description

To validate the viability of the utilized dataset, a deep analysis was performed to assess the diversity of the collected samples. To do so, the set of all day-separate intervals was evaluated in terms of protoform activation, confirming the presence or absence of all possible day patterns. As detailed in Section 2.1.2, protoforms follow the structure *R Q A are S*, combining five day moments, glucose labels, and four quantification types, resulting in 100 possible protoforms—of which nearly 90% were activated when processing the full dataset.

Only linguistic summaries covering a minimum of 75% of the maximum possible points for a day sampling are considered so, a total of

Table 7
Criteria description and associated evaluation method.

Fundamental	Criterion	Description	Evaluation
Information quality	Accuracy (Acc)	The provided response is correct and free of invalid statements.	The number of correct statements divided by the total of correct statements and the incorrect ones.
	Relevance (Re)	Output is free of superfluous and redundant information, omitting irrelevant measurements.	The intersection of human selected sentences and LLM selected sentences, divided by the human selected ones.
	Recall (R)	All important and necessary information is included.	The total of precise statements conditioned by all expected sentences.
	Correctness (Ct)	Incorrect statements are not contemplated along the output generation.	The number of correct statements (even if non-relevant) divided by all generated sentences.
	Reproductive (Rp)	Stability and control over the linguistic description over time, considering equal or similar queries.	Categorical: presence (1) or absence (0)
Thought capabilities	Understanding (Und)	LLM presents the capacity to fully understand the query, providing a meaningful, logical and contextual response.	Categorical: presence (1) or absence (0)
Communication quality	Clarity (C)	The linguistic description is clear and straightforward.	Categorical: presence (1) or absence (0)
	Accessibility (Ab)	The language used is not specialized in the domain, making the linguistic description understandable for every end user.	Categorical: presence (1) or absence (0)
	Temporal dynamics (Td)	Data description is made over time, referencing exact timestamps or concrete day intervals.	Categorical: presence (1) or absence (0)
Content safety	Hallucinations (H)	Non-description of invalid statements due to a lack of data understanding or interpretation.	Categorical: presence (1) or absence (0)
	Assumption verification (Av)	The response does not include interesting but non-existent facts that cannot be deduced from the input data.	Categorical: presence (1) or absence (0)

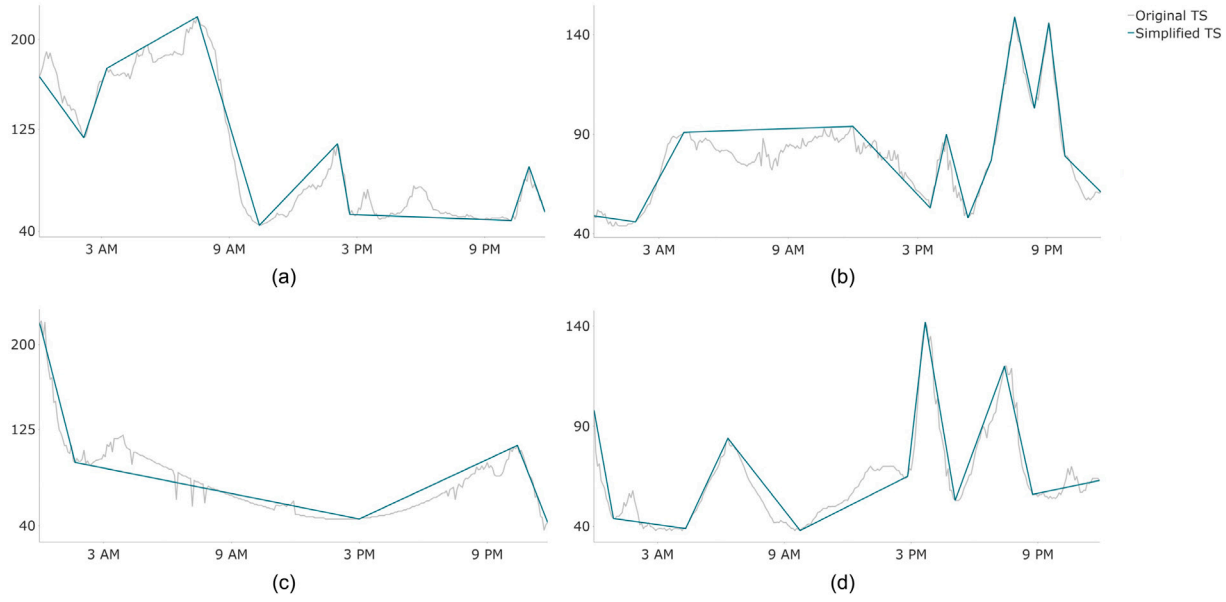

Fig. 3. Set of used TS for evaluation.

170 TS are included in the process divided into training (136 samples, 80%), validation (24 samples, 14.1%) and test sets (10 samples, 5.9%). Note that Algorithm 1 enabled the imputation of missing values in the selected TS, allowing us to establish that the dataset comprises approximately 48,960 CGM readings – assuming a maximum of 288 samples per day – and ultimately, the linguistic summary of each day.

Given the widespread popularization of GPT models in tasks including CGM data description [31], this manuscript presents the utilization of this LLMs series. It is highlighted that although GPT models are neither offered as open source nor free of charge when utilizing API services, the technological infrastructure required for training and running open-source models incurs a prohibitive cost, which may not be feasible to support the vast number of users anticipated for this kind of systems. In addition, this cost could also make it more difficult to replicate the methodology. Regarding this, our study focuses on fine-tuning ***gpt-3.5-turbo-1106***, ***gpt-4o-2024-08-06***, and ***gpt-4o-mini-2024-07-18***, the latest variants available [41]. Each was fine-tuned using either the complete glucose TS or a simplified segmented version, both linked to the same linguistic summary, resulting in six fine-tuned

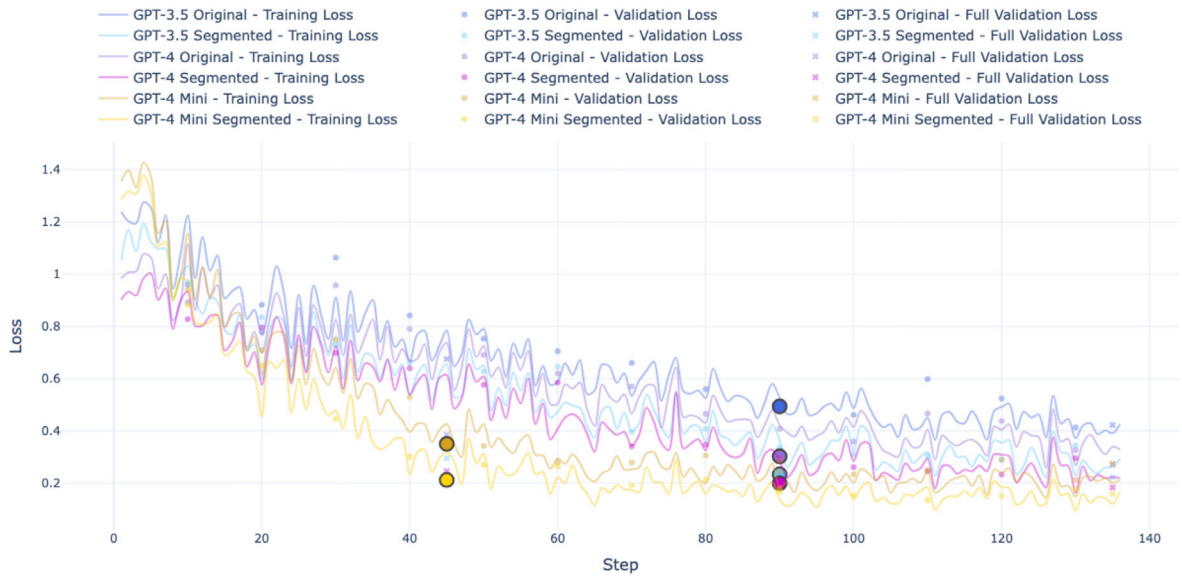

Fig. 4. Fine tuning process training loss along different models.

Table 8
Comparison of FT cost and configuration for different models. A summary.

Input data type	GPT-3.5		GPT-4o		GPT-4o Mini	
	Original TS	Segmented TS	Original TS	Segmented TS	Original TS	Segmented TS
Tokens	908,704	172,774	906,424	169,656	906,424	169,656
Training Cost	$7.27	$1.38	$22.66	$4.24	$1.36	$0.25
Query Cost	$0.0117	$0.0029	$0.0186	$0.0101	$0.0016	$0.0007

models and enabling evaluation of how data simplification impacts performance without altering semantic content.

3.2. Adaptation cycle of the LLMs

An iterative hyperparameter search was carried out to optimize the fine-tuned models, adjusting learning rate, batch size, and epochs based on training and validation loss trends. The final configuration — *learning rate multiplier = 0.8*, *batch size = 3*, *epochs = 3*, and *seed = 1 643 935 347* — was selected at the point where overfitting became evident, marked by stable training loss and a divergence in validation loss (see Fig. 4). Table 8 summarizes the final setup and associated costs for each model.

3.3. Outputs evaluation of the fine-tuned LLMs

Each TS (see Fig. 3) was evaluated using both the six fine-tuned models and the corresponding base ones, while prompt variations were also introduced to assess the impact of including domain and/or knowledge specifications. Table 9 presents the averaged results across the 4 TS for the original data, while Table 10 shows the results utilizing the segmented data. In total, 192 outputs are analyzed using the metrics outlined in Table 7. Table 11 highlights the best scores per model and the improvements obtained through PaP and FT. Lastly, Fig. 5 provides a visual overview of performance across executions based on *Acc*, *Re*, *R*, and *Ct*, with the remaining metrics supporting detailed analysis.

3.3.1. gpt-3.5 performance

For the original TS, *gpt-3.5* scores below 0.5 in *Acc* and *Re*, around 0.5 in *R* and *Ct*, with *Ct* reaching up to 0.8 in the base model. *Rp* only emerges in fine-tuned models with enriched prompts. *Und* improves with FT, especially when knowledge is included. *C* reaches 0.75 when being fine-tuned but drops with longer prompts. *Ab* is fully covered by the base model but drops with FT, scoring between 0.25 and 0.75. *H* is generally present except when prompts include domain and/or knowledge information after FT; *Av* behaves oppositely, dropping mainly when domain is added.

For segmented TS, *Acc* stays near 0.5, except when FT and domain information is present in the prompt. *Re* remains below 0.55 with low variance. *R* and *Ct* improve, especially in the base model with enriched prompts. *Rp* is always 0; *C* stays ≤ 0.5. *Ab* is mostly fulfilled. *Td* shows no clear pattern along the different configurations. *H* is 0 without prompt info but increases when knowledge (base model) or domain (FT) is added. *Av* scores 1 with FT, while fluctuating otherwise.

An output example is shown next for gpt-3.5:

- **Example:** 1
  - **Model:** gpt-3.5
  - **Fine-Tuning:** Yes
  - **Prompt:** [Header][Instruction][Knowledge]
  - **TS:** (a)
  - **Output:** *At the beginning of the day, glucose levels were high, with a maximum value of 190 at 00:25. Glucose levels decreased and reached a minimum value of 127.5 at 09:00, remaining* ***low*** *until 06:35* [...].

3.3.2. gpt-4o performance

In particular, when evaluating *gpt-4o* considering the original TS, the model achieves *Acc* and *Re* scores between 0.56 and 0.65, improving through FT, scoring up to 0.83, specially when knowledge information is available in the prompt. *R* and *Ct* are consistently high, with *Ct* reaching 1 across all FT configurations. *Rp* is 0 without FT and prompt information, but reaches up to 1 when both are provided. *C* ranges from 0.5 to 1, improving with simpler prompts. *Ab* scores 1 in all cases. *H* and *Av* are also mostly complete, except when the domain is included without FT, slightly reducing *Av*.

For segmented TS, *Acc* varies from 0.48 to 0.75, increasing with FT and knowledge; *Re* ranges from 0.55 to 0.69 with moderate variability.

Table 9
Evaluation metrics results for GPT models' performance utilizing the original TS for training.

Model	Tr	Dom	Kn	Acc	Re	R	Ct	Rp	Und	C	Ab	Td	H	Av	Score
3.5	✗	✗	✗	0.42 ±0.3	0.37 ±0.25	0.57 ±0.38	0.83 ±0.04	0	0.25	0.75	1	0.75	0	1	0.5
		✗	✓	0.44 ±0.08	0.37 ±0.15	0.54 ±0.1	0.63 ±0.09	0	0.5	0.25	1	0.75	0.25	1	0.52
		✓	✗	0.48 ±0.15	0.41 ±0.15	0.65 ±0.13	0.76 ±0.07	0	0.25	0.5	1	0.5	0	0.25	0.44
		✓	✓	0.36 ±0.17	0.34 ±0.16	0.5 ±0.27	0.55 ±0.31	0	0.5	0	1	0.75	0	1	0.46
	✓	✗	✗	0.34 ±0.14	0.31 ±0.18	0.44 ±0.25	0.47 ±0.3	0	0.25	0.75	0.75	0.5	0	1	0.39
		✗	✓	0.39 ±0.29	0.3 ±0.15	0.45 ±0.31	0.53 ±0.26	0.5	0.75	0.25	0.75	0.5	0	1	0.47
		✓	✗	0.38 ±0.22	0.33 ±0.18	0.52 ±0.34	0.64 ±0.24	1	0.5	0.75	0.25	0.75	0.75	0	0.49
		✓	✓	0.31 ±0.22	0.26 ±0.15	0.47 ±0.41	0.53 ±0.36	0.75	1	0.5	0.25	0.5	1	0	0.42
4o	✗	✗	✗	0.66 ±0.21	0.56 ±0.05	0.87 ±0.09	0.89 ±0.08	0	1	0.5	1	1	0.25	1	0.7
		✗	✓	0.6 ±0.14	0.58 ±0.08	0.91 ±0.1	0.94 ±0.07	0.75	0.5	1	1	1	1	1	0.84
		✓	✗	0.56 ±0.15	0.56 ±0.08	0.86 ±0.09	0.92 ±0.06	0.5	0.75	1	1	0.75	0.5	0.25	0.67
		✓	✓	0.6 ±0.15	0.64 ±0.11	0.89 ±0.14	0.91 ±0.11	0.25	0.75	0.75	1	1	0.75	0.75	0.75
	✓	✗	✗	0.61 ±0.12	0.64 ±0.17	1 ±0	1 ±0	0.5	1	1	1	1	1	1	0.87
		✗	✓	0.83 ±0.09	0.75 ±0.08	1 ±0	1 ±0	1	1	1	1	1	1	1	0.96
		✓	✗	0.62 ±0.13	0.62 ±0.09	0.95 ±0.1	0.97 ±0.06	0.5	0.75	0.75	1	1	1	1	0.83
		✓	✓	0.83 ±0.16	0.73 ±0.06	1 ±0	1 ±0	0.5	1	0.75	1	0.75	0.75	1	0.85
4o-mini	✗	✗	✗	0.51 ±0.11	0.44 ±0.04	0.78 ±0.16	0.83 ±0.14	0.75	0.75	0.5	1	0.5	0.75	0.75	0.67
		✗	✓	0.53 ±0.16	0.58 ±0.2	0.79 ±0.27	0.81 ±0.24	0.5	0.75	0.75	1	0.75	0.75	1	0.72
		✓	✗	0.45 ±0.17	0.44 ±0.12	0.64 ±0.14	0.76 ±0.1	0	0.25	0	1	0.25	0.75	0	0.37
		✓	✓	0.57 ±0.09	0.58 ±0.18	0.91 ±0.1	0.95 ±0.06	0	0.5	0.75	1	0.75	0.5	0.75	0.64
	✓	✗	✗	0.59 ±0.19	0.54 ±0.12	0.77 ±0.19	0.82 ±0.14	0.25	0.75	0.5	1	0.75	0.5	1	0.68
		✗	✓	0.63 ±0.16	0.48 ±0.14	0.85 ±0.18	0.85 ±0.18	0.75	0.5	0.75	1	1	0.5	1	0.76
		✓	✗	0.59 ±0.33	0.47 ±0.25	0.69 ±0.36	0.78 ±0.21	0.75	0.5	0.75	1	1	0.5	1	0.62
		✓	✓	0.51 ±0.06	0.42 ±0.11	0.6 ±0.11	0.65 ±0.12	0.25	0.5	0.75	0	0.75	1	0	0.56

R and *Ct* remain strong, between 0.76 and 1. *Rp* stays at 0 without FT, improving to 0.5 when FT or domain and/or knowledge are added. *Und* improves with simpler prompts. *C* fluctuates between 0.5 and 1, often higher when FT is not applied or knowledge is omitted. *Ab* and *Td* perform stably (0.75–1), especially with FT. *H* is full unless the prompt lacks knowledge after FT. *Av* is consistently high, further enhanced by FT.

These promising results have proven how gpt-4o performance is of high value as shown as follows on Examples 2 and 3 (output configuration correspond to the best setup specified in Table 11):

- **Example:** 2
  - **Model:** gpt-4o
  - **Fine-Tuning:** No
  - **Prompt:** [Header][Instruction][Knowledge]
  - **TS:** (a)
  - **Output:** *Throughout the day, the glucose levels exhibit notable fluctuations. In the early morning hours, from midnight to around 7 am, glucose levels are predominantly high, with several readings in the high range and peaking into very high levels around 7 am. This period is characterized by a peak reaching up to 219* [...].

- **Example:** 3
  - **Model:** gpt-4o
  - **Fine-Tuning:** Yes
  - **Prompt:** [Header][Instruction][Knowledge]
  - **TS:** (a)
  - **Output:** *At the beginning of the day glucose levels ere mostly high, with a hyperglycemia episode reaching very high values of 219 mg/dL. In the morning, there was a sharp decrease to low levels, followed by a hypoglycemia of 45 mg/dL. Later, in the morning, glucose increased to normal levels. At noon, glucose decreased again to low levels. In the afternoon, glucose experienced a valley at very low levels and remained low. [...] At the end of the day, glucose increased to normal levels and then decreased again to low levels.*

Table 10
Evaluation metrics results for GPT models' performance utilizing the segmented TS for training.

Model	Tr	Dom	Kn	Acc	Re	R	Ct	Rp	Und	C	Ab	Td	H	Av	Score
3.5	✗	✗	✗	0.52 ±0.15	0.38 ±0.14	0.64 ±0.18	0.69 ±0.2	0	0.5	0.25	0.5	0	0	0	0.36
		✗	✓	0.55 ±0.1	0.5 ±0.05	0.95 ±0.13	0.95 ±0.1	0	0.5	0.5	1	1	1	1	0.72
		✓	✗	0.51 ±0.11	0.53 ±0.14	0.89 ±0.21	0.93 ±0.14	0	0.75	0.5	1	0.75	0.75	0	0.6
		✓	✓	0.48 ±0.06	0.48 ±0.17	0.94 ±0.13	0.96 ±0.07	0	0.5	0.25	1	0.5	1	0.5	0.6
	✓	✗	✗	0.51 ±0.06	0.47 ±0.07	0.69 ±0.14	0.75 ±0.14	0	0.75	0.5	1	0.75	0.5	1	0.63
		✗	✓	0.54 ±0.11	0.44 ±0.17	0.65 ±0.27	0.66 ±0.27	0	0.5	0	1	0	0.5	1	0.48
		✓	✗	0.7 ±0.13	0.55 ±0.12	0.77 ±0.18	0.79 ±0.18	0	0.75	0.5	1	1	0.75	1	0.71
		✓	✓	0.56 ±0.23	0.49 ±0.19	0.67 ±0.25	0.73 ±0.23	0	0.25	0.5	1	0.5	0.25	1	0.54
4o	✗	✗	✗	0.64 ±0.09	0.68 ±0.09	1 ±0	1 ±0	0	1	0.75	1	1	1	1	0.82
		✗	✓	0.58 ±0.08	0.64 ±0.06	1 ±0	1 ±0	0.25	1	0.75	1	0.75	1	1	0.82
		✓	✗	0.48 ±0.15	0.55 ±0.08	0.93 ±0.13	0.98 ±0.05	0.5	1	0.75	0.75	1	1	0.75	0.77
		✓	✓	0.54 ±0.07	0.69 ±0.11	1 ±0	1 ±0	0	0.5	1	1	1	1	0.75	0.77
	✓	✗	✗	0.65 ±0.04	0.64 ±0.14	0.86 ±0.16	0.87 ±0.15	0.25	0.75	1	1	1	1	1	0.82
		✗	✓	0.75 ±0.04	0.59 ±0.11	0.87 ±0.08	0.88 ±0.08	0.5	0.75	0.5	1	1	0.5	1	0.76
		✓	✗	0.66 ±0.08	0.58 ±0.15	0.86 ±0.18	0.88 ±0.18	0.5	1	1	1	1	1	1	0.86
		✓	✓	0.7 ±0.22	0.56 ±0.27	0.76 ±0.22	0.88 ±0.18	0.25	0.5	0.5	1	1	0.5	1	0.69
4o-mini	✗	✗	✗	0.52 ±0.06	0.58 ±0.08	0.96 ±0.08	0.97 ±0.06	0	1	0.75	0.75	1	1	0.25	0.71
		✗	✓	0.51 ±0.05	0.67 ±0.1	0.89 ±0.13	0.93 ±0.1	0.25	0.5	0.5	1	1	1	1	0.75
		✓	✗	0.53 ±0.12	0.59 ±0.07	0.86 ±0.18	0.91 ±0.11	0	0.5	0.75	1	1	0.5	0	0.6
		✓	✓	0.5 ±0.06	0.59 ±0.07	0.89 ±0.13	0.92 ±0.09	0	0.75	0.5	1	1	0.5	0	0.6
	✓	✗	✗	0.65 ±0.09	0.51 ±0.12	0.74 ±0.12	0.76 ±0.13	0.25	0.75	0.5	1	1	0.75	1	0.72
		✗	✓	0.6 ±0.11	0.5 ±0.14	0.8 ±0.17	0.84 ±0.14	0.5	0.75	0.75	1	0.75	0.75	0.75	0.73
		✓	✗	0.65 ±0.15	0.52 ±0.1	0.86 ±0.19	0.87 ±0.19	0.25	0.75	0.5	1	1	0.75	1	0.74
		✓	✓	0.65 ±0.08	0.55 ±0.11	0.79 ±0.05	0.82 ±0.06	0.25	0.25	0.75	1	0.75	0.5	0.75	0.64

Table 11
Best performance for different configurations on input data.

Input data type	GPT-3.5		GPT-4o		GPT-4o Mini	
	Original TS	Segmented TS	Original TS	Segmented TS	Original TS	Segmented TS
Base model score	0.35	0.5	0.65	0.77	0.67	0.72
Best score (No FT)	0.52 (NoD-K)	0.72 (NoD-K)	0.84 (NoD-K)	0.82 (NoD-K)	0.72 (NoD-K)	0.75 (NoD-K)
Best score (FT)	0.49 (D-NoK)	0.71 (D-NoK)	**0.96** (NoD-K)	0.86 (D-NoK)	0.76 (NoD-K)	0.74 (D-NoK)

3.3.3. *gpt-4o-mini performance*

Focusing on the evaluation of *gpt-4o-mini* considering the original TS, the model shows variable *Acc* (0.45–0.63) and *Re* (0.42–0.58), with wide deviations, improving slightly with FT and enriched prompts. *R* and *Ct* remain consistently high (0.6–0.91 and 0.65–0.95). *Rp* improves only with FT. *Und* varies widely (0.25–1), depending on the provided prompt information. *C* scores range from 0.5 to 1, peaking with FT. *Ab* is generally stable, except when FT is combined with full prompt info. *Td* improves with FT. *H* is mostly high when the base model is utilized, registering lower score when FT is applied, except when both domain and knowledge are considered, where it increases to 1. Finally, *Av* experiences the opposite situation as *H*.

For segmented TS, *Acc* (0.5–0.65) and *Re* (0.5–0.58) show low variability, with minor improvements through FT. *R* and *Ct* remain high (0.74–0.96 and 0.76–0.97), performing better with the base model. *Rp* stays low, slightly improving with FT. *Und* is between 0.25 and 1, depending on the inclusion of domain or knowledge in the prompt. *C* fluctuates (0.5–0.75). *Ab* is stable (0.75–1), improving when knowledge is provided in the prompt. *Td* is mostly optimal (1), dropping only when knowledge is included in FT. *H* performs best in the base model. *Av* is more stable with FT, possibly dropping to 0 otherwise.

An output example of *gpt-4o-mini* is shown next:

- **Example:** 4
  - **Model:** gpt-4o-mini

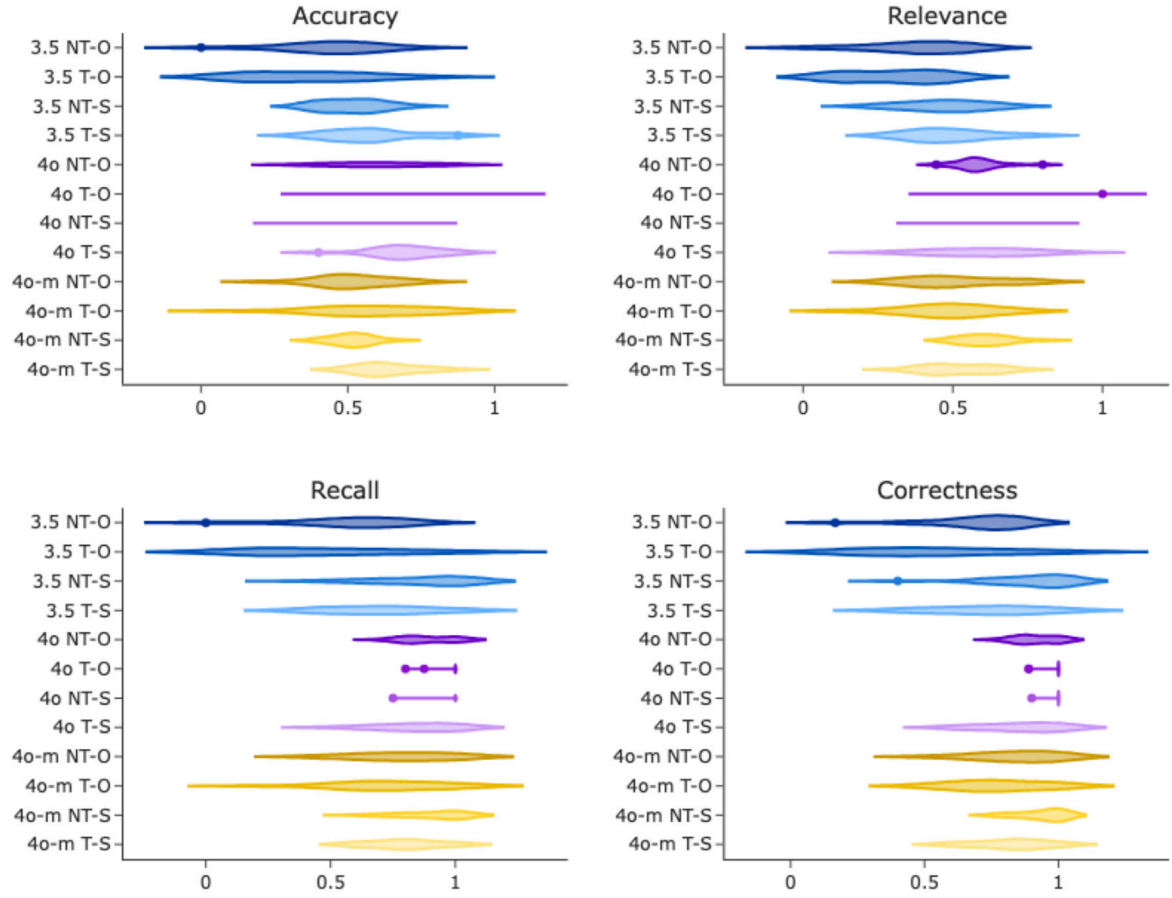

Fig. 5. In-depth analysis of information quality continuous metrics across different models.

- **Fine-Tuning:** Yes
- **Prompt:** [Header][Domain][Instruction] [Knowledge]
- **TS:** (c)
- **Output:** [...] *At the end of the morning, a peak of 62 mg/dL has been detected. In the afternoon, a peak of 60 mg/dL has been detected at around 11:55 am followed by a hypoglycemia episode with a minimum of 46 mg/dL for about 5 h. Later on, most glucose values have been low with an increasing trend during around 4 h. At the end of the day, many glucose values have been low with an increasing trend during around 2 h.*

4. Discussion

In this section, the discussion of the results obtained in Section 3 is exposed, aiming to further understand the performance of the evaluated models, their limitations, and the implications of the observed metrics in the context of the proposed methodology.

4.1. LLMs configuration

The hyperparameter search was conducted for all six models, but no significant improvements in loss trends were observed despite the optimization efforts. Consequently, a common configuration was applied uniformly across all models, ensuring a fair comparison. All models showed a decreasing training loss over time, indicating effective learning, despite some noise likely introduced by the stochastic optimization methods used by OpenAI.

These insights helped identify the checkpoints for the fine-tuned models. In all variants, validation loss began to increase after a certain point, suggesting that the models benefited from early stopping to prevent overfitting.

Table 8 outlines the final configuration and associated token costs. Notably, using segmented TS consistently reduced token usage and overall cost across models. The costs associated were particularly significant for gpt-4o when being trained with the original TS. Then, while segmentation improves efficiency, further analysis of output quality is essential to confirm that the models' performance is not compromised.

4.2. gpt-3.5 performance insights

In particular, the evaluation of *gpt-3.5* does not show clear improvements through the metrics utilized, since the model tends to include statements based on previous observations that are not present in the evaluated TS. This is reflected either using the original or the segmented TS. In fact, the less information provided to the model from the prompt, the better it performs, although there are no remarkable results in terms of accuracy and relevance as the metrics remain moderate. Nonetheless, the model has shown the capacity to avoid errors (Ct and R) when observing a segmented TS and only with the specification of the domain from the prompt (a general scenario that does not correspond to its best performance when only knowledge is provided). However, the results reflect the difficulty of the model to generate consistent outputs over time (Rp), plus the significant tendency to hallucinate when handling complex data, even though this was partially mitigated by segmenting the data, rather than FT.

4.3. gpt-4o performance insights

On its behalf, the results thrown by *gpt-4o* are remarkably strong across all aspects, due to its understanding (*Und*) capabilities for logical

reasoning as attested in [33], specially when the original TS is provided and the model has been fine-tuned. A particular behavior emerges throughout the executions: when the entire TS is contemplated and the model is fine-tuned, it shows fewer errors and achieves higher precision (Acc and Re). This setup proves outstanding, particularly when knowledge is specified in the prompt, representing the most promising configuration in this study. In contrast, when the TS is segmented, the model exhibits fewer errors when not being fine-tuned, displaying greater stability in these cases. Nonetheless, the fine-tuned *gpt-4o* with access to the original TS and explicit knowledge remains the top-performing setup in this case as well. Regarding the rest of the metrics, the model performs greatly in terms of clarity, using accessible language and secure statements, with a low trend towards hallucinations or statements that cannot be directly deduced from the input and available data. Nonetheless, it has been observed gpt-4o to provide redundant information when not being fine-tuned as shown in Example 2, while FT has allowed the model to aggregate similar sentences as exposed in Example 3. Taking all aspects together, it must be noted that *gpt-4o* has been conducted to avoid the misinterpretation of the task to perform, generate wrong clinical conclusions or hallucinate across data analysis as observed in [31]. At this point, it is highlighted how the summary provided in this work answers to the majority of questions defined in [32], unless those with the aim of performing a comparison between different days. In this way, we leverage the model to produce a meaningful and safe answer to be provided to the end-users without the need for further interaction with the model.

4.4. gpt-4o-mini performance insights

Focusing on *gpt-4o-mini*, it is appreciable how it falls between *gpt-3.5* and *gpt-4o*, showing little improvements in all metrics, though it does not reach the same level of consistency as *gpt-4o*. This model demonstrates high creativity, often including incorrect or irrelevant statements, consequently performing best when the TS is segmented and no FT is applied, as this configuration reduces the likelihood of errors. However, even in this scenario, there are clear signs of more correct or relevant information. Specifically, when the model is provided with a simple prompt, it tends to give a general description lacking specific details. On the other hand, when the domain is included in the prompt, the model starts to be more creative, often attempting o interpret data beyond what is explicitly available. In parallel, when knowledge is provided, the model attempts a step-by-step analysis, pretending an in-depth search within the TS, which enables fewer errors but bigger redundancies. As a result, as the complexity of the model increases in terms of training data and/or prompt length, the model in general starts to hallucinate, leading to a lack of clarity and logical consistency in the output.

4.5. Impact of prompt composition on results

In general, it has been shown that LLMs usually need knowledge specifications in the prompt when working with the original TS, regardless of whether they are fine-tuned or not. This avoids relying on implicit understanding, which can lead to irrelevant outputs or hallucinations. In contrast, when dealing with segmented TS, the base models still require explicit knowledge, but fine-tuned models perform well with domain information only. Although segmented approaches achieve acceptable results in terms of information quality (see Fig. 5), this seems to be mainly because the models capture simple patterns, without reaching deeper interpretations. Only *gpt-4o*, particularly when fine-tuned and using the original TS, shows the ability to adapt its output based on the prompt, as it adapts its performance based on the prompt it receives, avoiding major errors; while accuracy and relevance vary depending on the prompt's elements and converging towards the highest possible score with some prompt configurations, the recall and correctness remain consistently high. Finally, it is worth mentioning that models using segmented TS tend to struggle when summarizing, as they often fail to infer intermediate values or provide detailed descriptions.

4.6. Strengths, challenges and contributions

Finally, it has been observed that LLMs perform better on TS (a) and (d) from Fig. 3, compared to TS (b) and (c), which present prolonged periods of stability or consistent trends. In contrast, TS (a) and (d) contain more localized events. Since LLMs tend to follow sequential narratives, they often struggle to abstract or aggregate data over extended periods, rarely performing any meaningful quantification.

Although the methodology proposed here proves effective for summarizing CGM data on a daily basis, it also addresses concerns in the literature regarding the adaptation of LLMs to specific tasks [16,18–20,31]. The framework presented is potentially scalable beyond CGM data in diabetes. Moreover, the evaluation strategy adopted in this study helps overcome limitations identified in previous work, where model assessments often relied on subjective survey-based scoring without clear standards [25,31], or metrics that lack consistency across runs [32,35], also highlighting that general evaluation frameworks do not consider temporally dependent data [36]. In this sense, LLM-RawDMeth contributes to safe, accessible, and reproducible model evaluation by capturing variations through standardized metrics, supporting more objective and rigorous comparisons and ensuring reliable deployment in real contexts.

Future work should aim to scale the proposed methodology to other domains, assessing its robustness beyond the current context. Expanding the dataset and refining the fine-tuning process are also key to improving summary quality, particularly regarding information quality metrics. Ultimately, the goal is to support the summarization of longer time periods and more complex data, offering more personalized assistance to end-users and improving their quality of life.

5. Conclusions

This study presents a centralized IoT-driven architecture for CGM, enabling the generation of human-like linguistic summaries of glycemic dynamics through LLMs. The methodology involves expert-guided dataset generation, where each TS is associated with a linguistic summary, enabling FT of OpenAI models, in this case study, to optimize their performance. Prompt engineering experiments have also been conducted to identify configurations that best suit the specific characteristics of both the models and the available data. The comprehensive evaluation framework, LLM-RawDMeth, was developed with metrics that enable a thorough understanding of the model's behavior, providing insights into the model's performance and potential limitations, through the utilization of objective metrics to be mostly applied in domains with temporally dependent data.

The proposed system demonstrates the feasibility of daily summarization for CGM data, addressing the challenges in TS representation and natural language generation. Furthermore, the scalability of the evaluation framework and dataset methodology allows for broader applicability in domains requiring continuous monitoring and high-stakes decision-making.

CRediT authorship contribution statement

Juan F. Gaitán-Guerrero: Writing – original draft, Validation, Investigation. **Carmen Martínez-Cruz:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Macarena Espinilla:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization. **David Díaz-Jiménez:** Methodology, Conceptualization. **Jose L. López:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Statements of ethical approval

This study was reviewed and approved by the Ethics Committee of the University of Jaén (approval code: DIC.21/ 10.PRY). All procedures were conducted in accordance with international ethical standards, including the Declaration of Helsinki, and the participant provided written informed consent prior to their involvement in the study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This result has been partially supported by grant PID2021-127275OB-I00 and grant PID2021-126363NB-I00 funded by MICIU/AEI/10.13039/501100011033, Spain and by "ERDF A way of making Europe", and by grant PDC2023-145863-I00 funded by MICIU/AEI/ 10.13039/501100011033, Spain and by the "European Union NextGenerationEU/PRTR". Funding for open access charge: Universidad de Jaén/ CBUA

References

[1] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2023, http://dx.doi.org/10.48550/ARXIV.2303.18223.

[2] S. Altman, Planning for AGI and beyond, 2023, https://openai.com/index/planning-for-agi-and-beyond/. (Accessed 25 June 2024).

[3] K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A.T. Stüber, J. Topalis, T. Weber, P. Wesp, B.O. Sabel, J. Ricke, M. Ingrisch, ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports, Eur. Radiol. 34 (5) (2023) 2817–2825, http://dx.doi.org/10.1007/s00330-023-10213-1.

[4] J. Hu, T. Dong, L. Gang, H. Ma, P. Zou, X. Sun, D. Guo, X. Yang, M. Wang, Psycollm: Enhancing llm for psychological understanding and evaluation, IEEE Trans. Comput. Soc. Syst. (2024).

[5] O. Nov, N. Singh, D. Mann, Putting ChatGPT's medical advice to the (turing) test: Survey study, JMIR Med. Educ. 9 (2023) e46939, http://dx.doi.org/10.2196/46939.

[6] L. Campillos-Llanos, A. Valverde-Mateos, A. Capllonch-Carrión, Hybrid natural language processing tool for semantic annotation of medical texts in Spanish, BMC Bioinformatics 26 (1) (2025) 7.

[7] S. Chen, B.H. Kann, M.B. Foote, H.J. Aerts, G.K. Savova, R.H. Mak, D.S. Bitterman, The utility of ChatGPT for cancer treatment information, 2023, http://dx.doi.org/10.1101/2023.03.16.23287316.

[8] A. Shah, S. Chava, Zero is not hero yet: Benchmarking zero-shot performance of llms for financial tasks, SSRN Electron. J. (2023) http://dx.doi.org/10.2139/ssrn.4458613.

[9] H. Zhang, J.-J. Xu, H.-W. Cui, L. Li, Y. Yang, C.-S. Tang, N. Boers, When geoscience meets foundation models: Toward a general geoscience artificial intelligence system, IEEE Geosci. Remote. Sens. Mag. (2024).

[10] V. Sciannameo, D.J. Pagliari, S. Urru, P. Grimaldi, H. Ocagli, S. Ahsani-Nasab, R.I. Comoretto, D. Gregori, P. Berchialla, Information extraction from medical case reports using OpenAI InstructGPT, Comput. Methods Programs Biomed. 255 (2024) 108326.

[11] K. Rooney, OpenAI tops 400 million users despite DeepSeek's emergence, 2025, https://www.cnbc.com/2025/02/20/openai-tops-400-million-users-despite-deepseeks-emergence.html (Accessed 09 April 2025).

[12] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., ChatGPT for good? On opportunities and challenges of large language models for education, Learn. Individ. Differ. 103 (2023) 102274.

[13] A. Blair-Stanek, N. Holzenberger, B. Van Durme, Can GPT-3 perform statutory reasoning? in: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ACM, 2023, http://dx.doi.org/10.1145/3594536.3595163.

[14] J.H. Choi, K.E. Hickman, A. Monahan, D.B. Schwarcz, ChatGPT goes to law school, SSRN Electron. J. (2023) http://dx.doi.org/10.2139/ssrn.4335905.

[15] T. Feng, H. Shen, X. Yang, J.-M. Nianga, Z. Wang, Integration of large language models with IoT in smart agriculture to improve efficiency, yield, and quality, Ind. Sci. Eng. 1 (2024) 15–35.

[16] Z. Xu, Z. Wang, S. Li, X. Zhang, P. Lin, GeoPredict-LLM: Intelligent tunnel advanced geological prediction by reprogramming large language models, Intell. Geoengin. 1 (1) (2024) 49–57.

[17] X. Zhang, R.R. Chowdhury, R.K. Gupta, J. Shang, Large language models for time series: A survey, 2024, arXiv preprint arXiv:2402.01801.

[18] S. Ji, X. Zheng, C. Wu, HARGPT: Are LLMs zero-shot human activity recognizers? 2024, arXiv preprint arXiv:2403.02727.

[19] S.A. Imran, M.N.H. Khan, S. Biswas, B. Islam, LLaSA: Large multimodal agent for human activity analysis through wearable sensors, 2024, arXiv preprint arXiv:2406.14498.

[20] T. An, Y. Zhou, H. Zou, J. Yang, IoT-LLM: Enhancing real-world IoT task reasoning with large language models, 2024, arXiv preprint arXiv:2410.02429.

[21] R. Lou, K. Zhang, W. Yin, Large language model instruction following: A survey of progresses and challenges, Comput. Linguist. (2024) 1–10.

[22] M. Abbasian, Z. Yang, E. Khatibi, P. Zhang, N. Nagesh, I. Azimi, R. Jain, A.M. Rahmani, Knowledge-infused llm-powered conversational health agent: A case study for diabetes patients, in: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, IEEE, 2024, pp. 1–4.

[23] Z. Liu, C. Chen, J. Cao, M. Pan, J. Liu, N. Li, F. Miao, Y. Li, Large language models for cuffless blood pressure measurement from wearable biosignals, in: Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2024, pp. 1–11.

[24] H. Xu, L. Han, Q. Yang, M. Li, M. Srivastava, Penetrative ai: Making llms comprehend the physical world, in: Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications, 2024, pp. 1–7.

[25] C.M. Fang, V. Danry, N. Whitmore, A. Bao, A. Hutchison, C. Pierce, P. Maes, Physiollm: Supporting personalized health insights with wearables and large language models, 2024, arXiv preprint arXiv:2406.19283.

[26] International Diabetes Federation, Diabetes around the world in 2021, 2021, https://diabetesatlas.org/ (Accessed 25 June 2024).

[27] Dexcom, Dexcom continuous glucose monitoring, 2024, https://www.dexcom.com/ (Accessed 25 June 2024).

[28] Abbot, FreeStyle libre continuous glucose monitoring, 2024, https://www.freestyle.abbott/ (Accessed 25 June 2024).

[29] Abbot, FreeStyle libre 3 system, 2024, https://www.freestyle.abbott/us-en/products/freestyle-libre-3.html (Accessed 25 June 2024).

[30] D. Dao, J.Y.C. Teo, W. Wang, H.D. Nguyen, LLM-powered multimodal AI conversations for diabetes prevention, in: Proceedings of the 1st ACM Workshop on AI-Powered Q & A Systems for Multimedia, ICMR '24, ACM, 2024, http://dx.doi.org/10.1145/3643479.3662049.

[31] E. Healey, A.L.M. Tan, K.L. Flint, J.L. Ruiz, I. Kohane, A case study on using a large language model to analyze continuous glucose monitoring data, Sci. Rep. 15 (1) (2025) 1143, http://dx.doi.org/10.1038/s41598-024-84003-0.

[32] E. Healey, I. Kohane, LLM-CGM: A benchmark for large language model-enabled querying of continuous glucose monitoring data for conversational diabetes management, in: Biocomputing 2025: Proceedings of the Pacific Symposium, World Scientific, 2024, pp. 82–93.

[33] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Trans. Intell. Syst. Technol. 15 (3) (2024) 1–45.

[34] A.B. Sai, A.K. Mohankumar, M.M. Khapra, A survey of evaluation metrics used for NLG systems, ACM Comput. Surv. 55 (2) (2022) 1–39.

[35] E. Fons, R. Kaur, S. Palande, Z. Zeng, T. Balch, M. Veloso, S. Vyetrenko, Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark, 2024, arXiv preprint arXiv:2404.16563.

[36] T.Y.C. Tam, S. Sivarajkumar, S. Kapoor, A.V. Stolyar, K. Polanska, K.R. McCarthy, H. Osterhoudt, X. Wu, S. Visweswaran, S. Fu, et al., A framework for human evaluation of large language models in healthcare derived from literature review, NPJ Digit. Med. 7 (1) (2024) 258.

[37] Nightscout contributors, Nightscout xDrip+. URL https://github.com/NightscoutFoundation/xDrip.

[38] J.F. Gaitan-Guerrero, J.L. Lopez Ruiz, C. Martinez-Cruz, M. Espinilla, "T1GDUJA: Glucose dataset of a patient with type 1 diabetes mellitus", 2023, URL https://zenodo.org/records/10713570.

[39] L.A. Zadeh, Fuzzy logic, in: Granular, Fuzzy, and Soft Computing, Springer, 2023, pp. 19–49.

[40] N. Marín, D. Sánchez, On generating linguistic descriptions of time series, Fuzzy Sets and Systems 285 (2016) 6–30.

[41] OpenAI, Fine-tuning - OpenAI API, 2024, https://platform.openai.com/docs/guides/fine-tuning (Accessed 10 November 2024).

[42] T. Battelino, T. Danne, R.M. Bergenstal, S.A. Amiel, R. Beck, T. Biester, E. Bosi, B.A. Buckingham, W.T. Cefalu, K.L. Close, et al., Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range, Diabetes Care 42 (8) (2019) 1593–1603.

[43] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu, et al., Prompt engineering for healthcare: Methodologies and applications, 2023, arXiv preprint arXiv:2304.14670.

[44] OpenAI, Prompt engineering - OpenAI API, 2024, https://platform.openai.com/docs/guides/prompt-engineering (Accessed 25 June 2024).