

A Preliminary Study on NLP-Based Personalized Support for Type 1 Diabetes Management

Sandra Mitrović¹, Federico Fontana², Andrea Zignoli^{2,3}, Felipe Mattioni Maturana^{2,4}, Christian Berchtold¹, Daniele Malpetti¹, Sam Scott², Laura Azzimonti¹

¹Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), SUPSI, Switzerland

²Sestante Analytics AG, Switzerland

³Department of Industrial Engineering, University of Trento, Italy

⁴Department of Sports Medicine, University Hospital of Tübingen, Germany

Correspondence: sandra.mitrovic@supsi.ch

Abstract

The proliferation of wearable devices and sports monitoring apps has made tracking physical activity more accessible than ever. For individuals with Type 1 diabetes, regular exercise is essential for managing the condition, making personalized feedback particularly valuable. By leveraging data from physical activity sessions, NLP-generated messages can offer tailored guidance to help users optimize their workouts and make informed decisions. In this study, we assess several open-source pre-trained NLP models for this purpose. Contrary to expectations, our findings reveal that models fine-tuned on medical data or excelling in medical benchmarks do not necessarily produce high-quality messages.

1 Introduction

Type 1 diabetes (T1D) is an autoimmune disease in which the immune system attacks the pancreatic cells responsible for producing insulin, a hormone essential for regulating blood glucose levels. Without insulin, cells cannot absorb glucose, leading to potentially life-threatening consequences if not externally managed (NIH). Physical activity plays a crucial role in managing T1D, as it enhances insulin sensitivity, helps regulate blood glucose levels, and promotes overall health (Colberg et al., 2016). However, exercise must be carefully managed, as improper glucose regulation during physical activity can result in hypoglycemia (low blood glucose) or hyperglycemia (high blood glucose), both of which pose significant health risks.

Our study leverages data from wearable devices and insulin monitoring, together with domain expert inputs, to develop a Natural Language Processing (NLP)-based approach that generates personalized messages based on an individual's physical activity history. These messages, delivered after each activity session, help individuals adjust their behavior to minimize the risk of excessive glucose

fluctuations. This work serves as a proof of concept for the feasibility of using personalized NLP-driven messages in diabetes management, with the ultimate goal of driving behavior change.

In this article, we primarily focus on the message generation aspect of our work, evaluating several open-source models by assessing the quality of their generated messages and benchmarking them against expert-written text. The choice of open-source models was driven by our goal to integrate our approach into an app that prioritizes user privacy and transparency.

Related work Several studies have explored integrating wearable sensor data with NLP approaches, using data from various sources, including physiological and motion sensors (Ferrara, 2024), as well as smartphone, self-reported (e.g., survey responses) (Li et al., 2024a) and voice-assistant data (Li et al., 2024b). One notable example is PH-LLM (Cosentino et al., 2024), which introduces a fine-tuned large language model (LLM) based on Gemini, designed to interpret and reason over physiological data. Another example is PhysioLLM (Fang et al., 2024), which focuses on sleep monitoring by leveraging GPT-4-turbo. It enables users to analyze their sleep patterns and their impact on health, providing insights in the form of correlations and trends.

2 Problem and dataset description

Our aim was to generate short advice messages to be shown to app users at the end of a workout, to help them stay within normal glucose levels (70-180 mg/dL) during future workouts.

We broadly categorized workouts into two types: *negative-outcome* (i.e., the individual's glucose level went out of range, further subdivided into hypo- and hyperglycemia) and *positive-outcome* ones (i.e., the individual's glucose level remained within the range). *Negative-outcome* workouts re-

Outcome	Behavioral Pattern	Expert-provided Information	Ideal Response	Mistral Model	Hermes Model
negative (hypoglycemia)	You spent 18 minutes below 70 mg/dL during this workout.	Prolonged periods of hypoglycemia can lead to fatigue and potentially dangerous low glucose levels. Consider adjusting your carbohydrate intake before and during exercise.	To prevent hypoglycemia, consume 10–30 g of carbs before exercise if glucose is below 120–140 mg/dL, trending downward, or if active insulin is high. Adjust insulin in advance and monitor glucose regularly . Use fast-acting carbs during workouts if levels drop below 140 mg/dL.	To prevent prolonged hypoglycemia during exercise, consider extending your pre-workout carbohydrate intake and consume small, frequent carbohydrate snacks during your workout . Start your sessions later in the day to minimize the impact on your glucose levels.	In order to prevent hypoglycemia during your next workout, consider adjusting your carbohydrate intake prior to the session . As the start hour has no effect , focus on consuming carbs 18 minutes before the workout to help maintain normal glucose levels throughout the physical activity. Monitor your glucose levels closely and adjust your diet accordingly to ensure you stay within the normal range during exercise.

Table 1: An example of a hypo-glycemia scenario "Time spent <70mg/dL". To the right: models output for best actionable insights prompt with (0,0,1) (start time should increase). Color coding: black-action, red-hallucination, violet-nutrition (carbs), blue-insulin, green-monitoring, grey-specific quantity.

quire a behavior change, and messages should advise the individual on how to adjust. Conversely, if the *outcome* is positive, the individual should be encouraged to maintain their current behavior.

As such, we used a more detailed characterization of workouts, referred to as *behavioral patterns* (i.e., a more precise description of the out-of-range behavior selected from a pre-defined set of possibilities). Each behavioral pattern was accompanied by additional *expert-provided information*. We refer to the combination of *behavioral pattern* and *expert-provided information* as *observable insight*. An example of an *observable insight* is shown in the second and the third column of Table 1.

We considered a total of 17 different *observable-insight* scenarios: 13 with negative outcomes and 4 with positive outcomes. For 11 of the negative-outcome scenarios, we were also provided with the *ideal responses* generated by domain experts, which we used in our study as the ground truth.

In addition, we trained a machine learning model that also considered the activity history of the individual¹ to provide information on whether (and if so, how) changing their behavior with respect to at least one *actionable* variable could help the individual stay within the normal glucose range. Session intensity, duration, and start time were selected as the three *actionable* variables since they have easily interpretable meaning and are fully under the person’s control. We refer to the outcomes of the ML model as *actionable insights*.

¹Details about the model are not reported here due to space constraints and the focus on NLP methods.

3 Methodology

Model choice Given our commitment to using open-source models, we focused exclusively on these, excluding popular options like GPT-3.5 and GPT-4. Additionally, due to the limited size of the dataset available to us, fine-tuning was not a viable option, which further influenced our model selection process. The open-source models we considered are listed in Table 2.

Framework design We based our methodology on prompting (Brown et al., 2020) the open-source NLP models using different (types of) prompts in order to generate personalized user-friendly messages based on the aforementioned *insights*. Furthermore, as additional external knowledge from T1D experts (see *expert knowledge* in Table 1) was available, instead of relying solely on the pre-trained knowledge, we used that knowledge as additional information to enhance message generation. In other words, we simulated² Retrieval Augmented Generation (RAG) (Lewis et al., 2020) by extending our designed prompts with the related expert knowledge as the relevant *context*.

Prompt types We designed two types of prompts: one that receives only *observable insights*, and another that on top of these includes *actionable insights*. In the first iteration, we evaluated a wide range of models using the *observable-insight*

²We opted for not implementing a separate retrieval component for a RAG system as the additional inputs are too short and our dataset is too small to justify it.

Model (Owner)	Motivation
Starling-LM-7B-beta (Nexusflow)	FT; OML
gemma-2-2b-it (Google)	instruct FT; OML
Mistral-7B-Instruct-v0.3 (MistralAI)	instruct FT; OML
Hermes-2-Pro-Mistral-7B (NousResearch)	FT; OML
JSL-MedPhi2-2.7B (John Snow Labs)	FT on medical data
Llama-3.2-3B-Instruct (Meta)	instruct FT
falcon-7b-instruct (TII)	chat/instruct FT

Table 2: List of the open-source models and motivation for choosing them. "FT" stands for fine-tuned, while "OML" denotes the model's high-rank (among top-4) at The Open Medical-LLM Leaderboard (Pal et al., 2024).

prompt type. The rationale was that models performing poorly on simpler prompts should not be considered for more complex ones. Following the initial selection, we introduced the second type of prompt (*full-insight*), which also includes *actionable insights*. For example, in the scenario from Table 1, the desirable behavior might involve providing suggestions on how the user should adjust their activity to avoid hypoglycemia, such as reducing workout intensity. We encoded actionable insights from our ML model as a three-dimensional vector, with dimensions corresponding to session intensity, duration, and start hour, respectively. Each dimension can take one of three values (0, -1, 1), where 0 denotes no effect on glucose level (hence no action required), -1 denotes that the variable should be decreased to reduce out-of-range risk, while +1 denotes that the variable should be increased for the same purpose. A well-performing model should be able to incorporate this information and generate a corresponding message. Both *observable-* and *full-insight* prompts were then iteratively refined based on the evaluation criteria described below.

Evaluation We evaluated the generated messages both qualitatively and quantitatively. The qualitative analysis focused on prompt adherence, correctness, level of detail, emotional tone, and medical content comprehension. Quantitative analysis was feasible only for *observable-insight* prompts, where ground truth allowed comparison by measuring semantic similarity between generated messages and expert-provided ideal responses.

4 Results

Qualitative analysis The formulation of the initial *observable-insight* prompt (P) can be seen in Figure 1. We found that many models have difficulty adhering to this prompt. Besides *Gemma*, that

Initial *observable-insight* prompt (P): "You will be provided with an observed PATTERN in a physical activity session of a person with a T1 diabetes condition. You are supposed to generate a 15-20 words long ADVICE related to the observed pattern that can help the person to stay within normal glucose levels. You should also incorporate the given CONTEXT."

Best *full-insight* prompt: "You will be provided with an observed PATTERN in a physical activity session of a person with a T1 diabetes condition which leads to a particular EFFECT in person's glucose level. The PATTERN is a three-dimensional vector where the first component refers to session intensity, the second component refers to session duration and the third component refers to session's start hour. Each of the three components can take exactly one of the values -1, 0 or 1. Value 0 denotes that i-th component has no effect at all on person's glucose level and as such it is not relevant for the advice. Non-zero value on i-th position denotes that the i-th component has negative effect on person's glucose level and is very relevant for the advice. Value -1 on i-th position denotes that decrease in the i-th component would make person stay within the normal glucose levels. Value 1 on i-th position instead denotes that increase in the i-th component would make person stay within the normal glucose levels. You are supposed to generate a 15-20 words long ADVICE leveraging the observed PATTERN and the given CONTEXT to help the person stay within normal glucose levels, commenting only on those particular components that have an effect on the patient glucose levels in the given case. Make sure to take into account glucose level when giving the advice and not to mix recommendations for hypoglycemia (glucose level below 55 mg/dL) with those for hyperglycemia (glucose level above 180 mg/dL). Keep a professional tone and focus more on prevention that can be done prior and/or during the workout instead of giving a precise medical advice. PATTERN: (X,Y,Z)."

Figure 1: Examples of prompts. In the second one, (X,Y,Z) is the 3-dimensional output of the ML model. Moreover, the underlined parts are shared with the best-performing *observable-insight* prompts.

had different problems (see examples 1 & 2 in Table 3), *Starling-LM* and *JSL-MedPhi2* consistently generated overly long responses (here omitted in the interest of space). The same issue was often also observed with *Falcon*. Even more importantly, *JSL-MedPhi2* frequently exhibited undesired behavior: after a paragraph it would either start repeating itself or encounter hallucination issues.

Examples 3-5 in Table 3 illustrate other issues which do not necessarily stem from not adhering to the prompt, but nevertheless were judged by experts as inappropriate. We, therefore, implemented an iterative-refinement strategy in order to: 1) reduce emotional expressiveness (denoted as: P_e) and avoid giving medical advices (P_m); 2) focus on the observed behavioral pattern (P_f); 3) enforce

Ex.	Scenario	Model	Response	Problem/Experts comment
1	Time spent below 55 mg/dL	Gemma	"Increase carbohydrate intake before and during exercise to avoid prolonged hypoglycemia."	Not adhering to prompt (length)
2	Time spent above 250 mg/dL (if workout >2h in duration)	Gemma	"Monitor your glucose closely and adjust insulin as needed to stay within target range."	Lack of necessary detail, too general and ignoring context (expert input)
3	Time spent above 250 mg/dL (if workout >2h in duration)	Llama	"Reassess insulin strategy and carbohydrate intake to prevent glucose spikes, aiming for a 5-10% increase in daily insulin doses."	Wrong type of response: providing medical advice instead of general well-being recommendations
4	Time spent below 55 mg/dL	Mistral Llama Gemma	"Immediately consume fast-acting..." "Consume ... within the next 5 minutes" "Consume ... carbs immediately"	Time aspect ("immediately" or "within the next 5 minutes")
5	Glucose variability <36% during exercise	Gemma	"Keep up the good work! Your consistent glucose management is key to optimal performance."	A bit too much excessive emotional expressiveness.

Table 3: Examples illustrating some of the challenges with generated responses using the initial *observable-insight* prompt (P) with different NLP models.

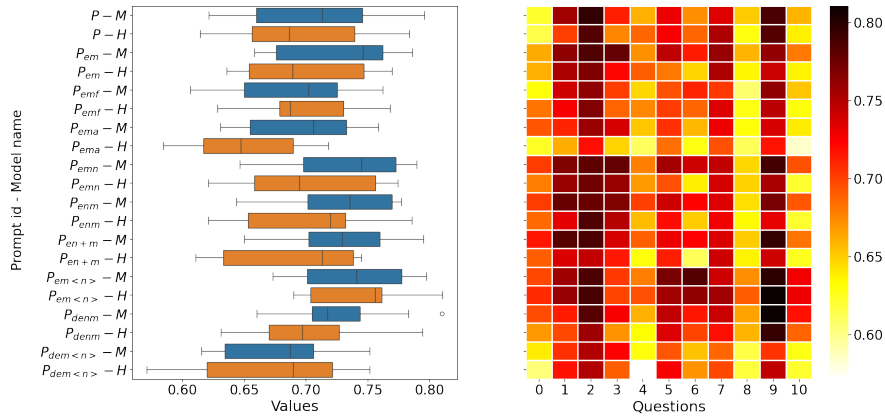


Figure 2: Similarity scores for each *observable-insight* prompt and model combination. To the left: boxplots per prompt-model combination over 11 questions, to the right: heatmap with a refined view on a question level.

generation of concrete actions instead of vague suggestions (P_a); 4) incorporate sports nutrition guidelines (short form: P_n , more elaborate: P_{n+} , explicitly adding a summary of sports nutrition guidelines to the context: $P_{<n>}$); 5) differentiate between hypo-/hyper-glycemia in responses (P_d).

Based on the expert feedback, we shortlisted *Mistral* and *Hermes* (which uses a previous version of *Mistral* as a base model) to conclude our experiments and explore *full-insight* prompts. Among these, the best performance was obtained with the prompt shown in Figure 1. This prompt works well with the *Mistral* model, while *Hermes* produces more hallucinations (see rightmost part of Table 1).

Quantitative analysis For *observable-insight* prompts, we evaluated the semantic similarity scores between generated and ideal responses using *all-MiniLM-L6-v2* (Reimers and Gurevych, 2019). The left panel of Figure 2 shows boxplots of similarity scores for each prompt-model combination, with *Mistral* demonstrating superior performance

in the majority of cases. To provide a more granular view, the right panel of Figure 2 presents a heatmap of similarity scores for each prompt-model combination and for each question. Notably, for some questions it is difficult to reach high similarity level regardless of the model+prompt choice (e.g., question 8). Additionally, the superior performance of *Mistral* is again evident, as seen, e.g., in the clear horizontal patterns for questions 4, 6, and 10.

5 Conclusions

This work evaluates the feasibility of generating high-quality, personalized NLP-based messages for diabetes management, integrating both domain expert inputs and data-driven insights, with the goal of driving behavior change. We tested several open-source models, and among them, *Mistral* yielded particularly promising results, proving to be a strong candidate for this task. In contrast, *Starling-LM-7B-beta*, *gemma-2-2b-it*, *Llama-3.2-3B-Instruct*, and especially, *JSL-MedPhi2-2.7B* delivered disappointing outcomes.

Limitations

We acknowledge that the limited size of our dataset affects the generalizability of our conclusions. However, we hope our findings offer valuable insights, particularly by encouraging caution when using models we found less effective for similar tasks.

Additionally, we attempted to improve the safety and quality of generated messages, in particular by introducing prompt refinement strategies $P_m, P_f, P_a, P_n, P_{n+}, P_{<n>}, P_d$. However, we recognize that further investigation is needed to fully address this issue.

Moreover, as a proof of concept, our approach considered only a limited set of actionable variables, and its scalability to a broader set should be explored in future work.

Ethical Considerations

We strongly advise against any use of suggested prompts that breaches ethical standards or facilitates harmful activities, such as generating misleading, harmful or malicious content. Our commitment to ethical principles underscores our dedication to fostering a positive impact not only in the related research community but, even more importantly, in all related fields and domains where potential real world utility and applicability of this work exists.

Acknowledgements

The work described in this paper has been supported by Innosuisse - Swiss Innovation Agency, project no. 72257.1, INNO-ICT.

References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Sheri R Colberg, Ronald J Sigal, Jane E Yardley, Michael C Riddell, David W Dunstan, Paddy C Dempsey, Edward S Horton, Kristin Castorino, and Deborah F Tate. 2016. Physical activity/exercise and diabetes: a position statement of the american diabetes association. *Diabetes care*, 39(11):2065.
- Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. 2024. Towards a personal health large language model. *arXiv preprint arXiv:2406.06474*.
- Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. 2024. Physiollm: Supporting personalized health insights with wearables and large language models. *arXiv preprint arXiv:2406.19283*.
- Emilio Ferrara. 2024. Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: a survey of early trends, datasets, and challenges. *Sensors*, 24(15):5045.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Jiachen Li, Justin Steinberg, Xiwen Li, Akshat Choube, Bingsheng Yao, Dakuo Wang, Elizabeth Mynatt, and Varun Mishra. 2024a. Vital insight: Assisting experts’ sensemaking process of multi-modal personal tracking data using visualization and llm. *arXiv preprint arXiv:2410.14879*.
- Jiachen Li, Justin Steinberg, Xiwen Li, Bingsheng Yao, Dakuo Wang, Elizabeth Mynatt, and Varun Mishra. 2024b. Understanding the daily lives of older adults: Integrating multi-modal personal health tracking data through visualization and large language models. In *Proceedings of the AAAI Symposium Series*, volume 4, pages 173–177.
- NIH. Type 1 Diabetes - NIDDK — niddk.nih.gov. <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-1-diabetes>. [Accessed 12-12-2024].
- Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, and Beatrice Alex. 2024. openlifescienceai/open_medical_llm_leaderboard. https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard.
- Nils Reimers and Iryna Gurevych. 2019. *Sentence-bert: Sentence embeddings using siamese bert-networks*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.