What do LLMs value? An evaluation framework for revealing subjective trade-offs in assessment of glycemic control

Payal Chandak* CHANDAK@MIT.EDU
Harvard-MIT Program in Health Sciences and Technology

Elizabeth Healey* ELIZABETH.HEALEY@CHILDRENS.HARVARD.EDU
Boston Children's Hospital

Maria F. Villa-Tamayo MZU4GE@VIRGINIA.EDU
University of Virginia

Agatha F. Scheideman SCHEIDEMAN@DIABETESTECHNOLOGY.ORG
Diabetes Technology Society

Mandy M. Shao SHAO@DIABETESTECHNOLOGY.ORG
Diabetes Technology Society

Chiara Fabris CF9QE@VIRGINIA.EDU
University of Virginia

Kenneth D. Mandl KENNETH.MANDL@CHILDRENS.HARVARD.EDU
Boston Children's Hospital

Isaac Kohane ISAAC_KOHANE@HMS.HARVARD.EDU
Harvard Medical School

David C. Klonoff DKLONOFF@DIABETESTECHNOLOGY.ORG
Diabetes Research Institute, Mills-Peninsula Medical Center (Sutter Health)

Abstract

Clinical decisions often require balancing conflicting priorities rather than simply selecting a single "correct" answer. We present an evaluation framework that probes the value judgments embedded in large language models (LLMs) by testing how they assess quality of glycemic control from continuous glucose monitoring (CGM) data. Using synthetic type 1 diabetes profiles, we asked five commercial LLMs to perform pairwise comparisons of CGM summary statistics and derived a percentile ranking for each profile. We then quantified alignment with two reference metrics: time in range (TIR) and the expert-derived Glycemia Risk Index (GRI), which was developed with clinician input regarding preferences across glycemic ranges. Across three insulin therapy modalities, newer models showed stronger correlation with GRI than older models, suggesting a generational shift toward expert consensus. However, a perturbation analysis revealed instances of disagreement around the weighting of mild hypoglycemia and mild hyperglycemia relative to the GRI. These results demonstrate that high average agreement with clinical metrics can mask clinically meaningful misalignments in how LLMs prioritize risks. Our proposed framework reveals how LLM outputs reflect competing priorities in clinical contexts.

Keywords: large language models, diabetes, explainable artificial intelligence

Data and Code Availability We simulated patient data using the UVA/Padova T1D FDA-accepted patient simulator (Man et al., 2014). We have included code at https://github.com/lizhealey/LLMGRI.

Institutional Review Board IRB approval was not necessary since this study used synthetic data.

* These authors contributed equally (alphabetical)

© 2025 P. Chandak, E. Healey, M.F. Villa-Tamayo, A.F. Scheideman, M.M. Shao, C. Fabris, K.D. Mandl, I. Kohane & D.C. Klonoff.

1. Introduction

Diabetes care is ripe for the deployment of LLMs (Pavon et al. (2025)), with recent work investigating their potential to assist in the summarization of glucose data (Cardei et al. (2025); Healey et al. (2025); Choi and Raj (2025)). Despite their potential, LLMs have seen limited adoption in medical decision-making scenarios, especially those that demand the balancing of competing risks and interpretation, such as in diabetes management. While randomized controlled trials (RCTs) (The Diabetes Control and Complications Trial Research Group, 1993; ADVANCE Collaborative Group et al., 2008) have demonstrated a clear benefit of intensive insulin therapy to achieve glycemic control to avoid long-term diabetes complications, intensive insulin therapy comes with the risk of hypoglycemia, which has its own acute and long-term consequences (McCall, 2012). Thus, clinicians frequently make clinical judgments when adjusting insulin regimens to achieve targets for glycemic control and while guidelines exist (ElSayed et al., 2022), there are information gaps around precise recommendations for glycemic targets.

This uncertainty is not limited to diabetes, as many areas of clinical decision-making lack evidence (Frieden, 2017; Bothwell et al., 2016; Djulbegovic and Guyatt, 2019). Both surgeons and psychiatrists have argued that it is inappropriate, and sometimes impossible, to evaluate interventions with RCTs (Bonchek, 1979; Ablon and Jones, 2002). Even interventions that are backed by RCTs require weighing benefits against potential harms. For example, intensive blood pressure lowering reduces the risk of mortality and cardiovascular events but increases the risk of acute kidney injury (SPRINT Research Group et al., 2015). In high-stakes scenarios with competing medical risks, clinical decision making is a balancing act between evidence-based medicine and consensus-based practice that requires constant value judgements (Frieden, 2017; Bothwell et al., 2016; Djulbegovic and Guyatt, 2019). As LLMs become increasingly integrated into clinical workflows, an important question arises: do these models understand and reflect the reasoning and expert medical judgment of clinicians (Andrew Taylor, 2025)?

In diabetes management, clinicians frequently make judgments when adjusting insulin regimens to achieve targets for glycemic control. Clinician subjectivity is especially pronounced in interpretation of glycemic control derived from CGM data, which gives comprehensive insight into past glycemic control. In making insulin dosing recommendations, clinicians must weigh the chronic, long-term risks of hyperglycemia against the acute, potentially life-threatening dangers of hypoglycemia (ElSayed et al., 2022). A long-standing indicator of glycemic quality derived from CGM, the percentage of time that glucose remains in range (TIR) (Dovc and Battelino, 2021) has been criticized for its insensitivity to these critical tradeoffs (Rodbard, 2021). In real-world practice, interpretation of glycemic control often differs significantly across clinicians (Nimri et al., 2018; Spartano et al., 2025). This reflects differing priorities, risk tolerances, and value frameworks. However, it remains unknown how LLMs make qualitative assessments about glycemic control.

Although substantial portions of medical practice are consensus-based, there are few evaluation methods specifically assessing whether LLMs reflect judgements of expert clinicians (Yu et al., 2024). Traditional benchmarks tend to focus on clinical decisions that are grounded in evidence, prioritizing factual accuracy and guideline adherence. Evaluation frameworks often struggle with navigating a spectrum of subjective tradeoffs because they assume that "correctness" can always be objectively benchmarked against an established gold standard (Raji et al., 2025). However, in contexts where clinical judgement requires weighing competing harms, like diabetes management, this assumption breaks down, and evaluating LLM alignment becomes more complex.

The challenge of consensus-based practice is further compounded by the difficulty of articulating clinical value frameworks. When assessing glycemic control, clinicians consider multiple contextual factors, including the time spent in hypo– and hyperglycemia and the severity of these glycemic excursions. These judgments are often tacit and situational, making them hard to quantify or encode into traditional evaluation pipelines. To fill the gap for a consensus–based measure of glycemic control, the Glycemia Risk Index (GRI) was recently introduced as a metric grounded in endocrinologists' rankings of CGM profiles (Klonoff et al., 2023). The GRI assigns different weights to the contextual factors of time spent in hypoglycemia and hyperglycemia. This weighting approximates the implicit values that clinicians bring to their assessments of glycemic control. This risk score presents an opportunity for evaluating how well LLMs are aligned with the subjective tradeoffs

when assessing the quality of glycemic control from CGM data.

In this work, we introduce a framework for evaluating whether LLMs align with the implicit judgements of clinical experts. Using the concrete example of glycemic control in the setting of type 1 diabetes (T1D), we investigate whether LLM preferences align more closely with the heuristics of TIR, or the expert-derived metric, the GRI. We compare five frontier models from two commercial creators (GPT 3.5, GPT 4.1, o4-mini, Claude 3.7 Sonnet, and Claude 3 Haiku) against both TIR and GRI benchmarks. Notably, GPT 3.5 and Claude 3.7 Sonnet were developed prior to the introduction of the GRI, while the other models were developed after, which we hypothesize may have an effect on how the LLMs synthesize nuances of time spent in hypoglycemia and time spent in hyperglycemia. Through a series of perturbation experiments, we find areas of disagreement between LLMs and expert-derived metrics, particularly around weighting of mild hypoglycemia and mild hyperglycemia. Through our explainable artificial intelligence (AI) approach, we highlight the need for evaluation methods that move beyond binary correctness and toward alignment with clinical reasoning. By grounding LLM assessment in real-world clinical tradeoffs, our approach contributes to the broader goal of value-sensitive AI in medicine. As LLMs are increasingly tasked with supporting clinical decisions, ensuring that they reason in ways that are interpretable, trustworthy, and aligned with expert judgment will be essential for their future deployment.

2. Background

2.1. Glycemic control

When reviewing glycemic profiles, clinicians typically refer to an Ambulatory Glucose Profile report (AGP) (International Diabetes Center). This report presents 14 days of CGM data. It consists of a 24-hour mean glucose profile (also called the AGP), daily profiles, and a section comprised of seven key metrics: mean glucose, represented both by an average glucose concentration and another number proportional to the mean glucose called the Glucose Management Indicator (GMI); coefficient of variation; time in range (TIR) (70 mg/dL – 180 mg/dL); time below range 1 ($\geq$ 54 mg/dL, $<$ 70 mg/dL); time below range 2 ($<$ 54 mg/dL); time above range 1 ($>$ 180 mg/dL, $\leq$ 250 mg/dL); and time above range 2 ($>$ 250 mg/dL) (Battelino et al., 2019).

Of these metrics, TIR has long been viewed as a useful single metric to assess control because of its correlation to HbA1c and diabetes outcomes (Dovc and Battelino, 2021; Aleppo, 2021) . However, international guidelines on interpretation note that TIR should be interpreted in the context of time spent in hypoglycemia and time spent in hyperglycemia (Battelino et al., 2019).

2.2. The Glycemia Risk Index

The Glycemia Risk Index (GRI) (Klonoff et al., 2023) was developed in 2023 to serve as a single composite metric for assessment of glycemic control. It was created through the input of clinicians and was designed to align with how clinicians assess glycemic control, given the metrics found on the AGP report and the 24-hour glucose profile. Instead of leveraging the seven metrics on the AGP individually, the GRI accounts for all seven metrics weighted according to preferences of 330 experienced endocrinologists and calculated from a formula that incorporates percentages spent in the four glucose ranges outside of TIR. The GRI, as presented in Equation 1, is a linear combination of percentages with a GRI of 0 corresponding to perfect control (100% TIR), and a higher score indicating worse control, with the most significant weight on severe hypoglycemia. Klonoff et al. found that the GRI was more closely aligned to the clinician percentile ranking of CGM profiles than other commonly used metrics, such as TIR.

$$\begin{aligned}\text{GRI} =&(3.0 \times \text{Severe Hypo}) + (2.4 \times \text{Hypo})\\ &+(1.6 \times \text{Severe Hyper}) + (0.8 \times \text{Hyper}) \quad (1)\end{aligned}$$

where Severe Hypo, Hypo, Hyper, and Severe Hyper correspond respectively to the percentage of time spent in the ranges $<$ 54 mg/dL, $\geq$ 54 and $<$ 70 mg/dL, $>$ 180 and $\leq$ 250 mg/dL, and $>$ 250 mg/dL.

The development of this risk index has positive implications for how glycemic control can be assessed, as the availability of a single, quantitative metric for describing the quality of glycemic control can be useful in assessing how glycemic control changes over time. In our setting of distilling values embedded in the LLMs, the GRI can be useful because it serves as a proxy for how clinicians assess the quality of glycemic control. Although the fit of GRI to the clinician-ranked percentile of CGM traces was not perfect (ad-

justed $R^2 = 0.904$) (Klonoff et al., 2023), given the expected variability in clinician assessment, we use it as a quantifiable metric of glycemic control.

3. Methods

Data. We simulated patient data using the UVA/Padova T1D FDA-accepted patient simulator (Visentin et al., 2018). We chose to use simulated data for multiple reasons. The patient simulator had 100 unique metabolic profiles of adult subjects, allowing us to work with a large cohort of complete CGM profiles. By using simulated CGM data, we were also able to generate a range of glycemic profiles with different levels of control. Additionally, the synthetic data allowed us to leverage commercial language models without restrictions for patient privacy.

The simulator supports multiple treatment modalities. We generated three separate datasets corresponding to the following insulin therapies:

1. **MDI**: Multiple daily injections regimen simulated with a daily long-acting insulin basal dose and fast-acting insulin boluses provided at mealtimes.

2. **CSII**: Continuous subcutaneous insulin infusion (open-loop pump therapy) simulated using individualized therapy profiles.

3. **CLC**: Closed-loop control therapy simulated using the USS-Virginia algorithm, the academic version of a commercially available automated insulin delivery system (Control-IQ).

See Appendix A for detailed descriptions of the dataset generation.

Figure 1: Assessing LLM alignment to the GRI. Each dataset had 4950 pairwise comparisons of every profile. Each of the 100 profiles was given a rank based on the number of other profiles the LLM decided that it was better than. This percentile rank was then compared to the GRI.

LLM Alignment with the GRI. The first set of experiments was designed to compare the preferences embedded in LLMs to a known clinical risk index, the GRI. This was done by having the LLM perform pairwise comparisons to generate a percentile ranking of glycemic profiles. Figure 1 shows the overview of the LLM-ranking experiments. Our evaluation used five commercial LLMs from two companies (Table 2). We used three OpenAI GPT series models (OpenAI et al., 2023; Brown et al., 2020) and two Anthropic Claude series models (Anthropic, 2025b). Figure 2 shows the prompt design. See Appendix B for additional details on the setup.

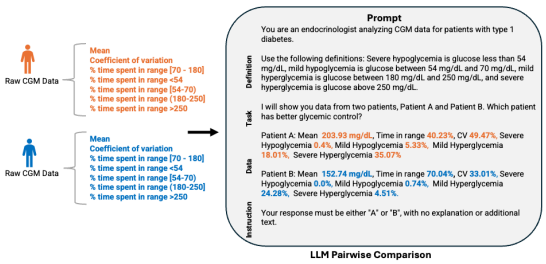

Figure 2: Prompt for a given pairwise comparison.

Perturbation Analysis. To examine how LLMs balance hypoglycemia and hyperglycemia, we generated additional synthetic glycemic profiles with systematically varied time in different glycemic ranges. Models were prompted to make pairwise comparisons of which profiles were "better," benchmarked against 100 individuals from the CSII dataset. Two experiments were performed: (1) varying mild hypoglycemia vs. mild hyperglycemia, and (2) varying severe hypoglycemia vs. severe hyperglycemia. An overview is shown in Figure 7 and details are provided in Appendix C.

Regression analysis. We used the LLMs percentile ranks to perform a regression analysis on the four variables used in the GRI formula: % Severe Hypo, % Mild Hypo, % Mild Hyper, and % Severe Hyper. For the perturbation analysis, we used the ranking against the profiles in the CSII dataset and the percentage time spent in hypoglycemia and hyperglycemia. The mild hypoglycemia and mild hyperglycemia perturbation analysis used only the percent time spent in mild hypoglycemia and mild hyperglycemia in the regression. The severe hypoglycemia and severe hyperglycemia perturba-

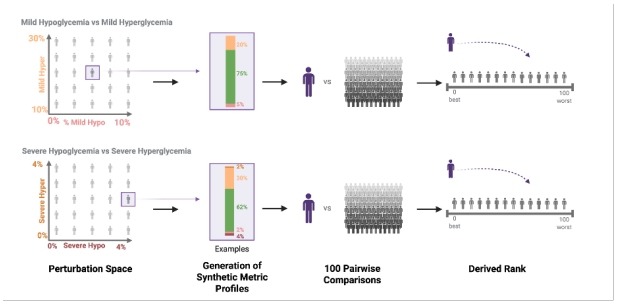

Figure 3: Overview of perturbation analysis experiments. Synthetic metric profiles were generated based on different compositions of time in hyperglycemia and hypoglycemia. Top: perturbations of mild hypoglycemia and mild hyperglycemia with severe metrics fixed at 0%. Bottom: perturbations of severe hypoglycemia and severe hyperglycemia with mild metrics fixed at 2% and 30%, respectively. Each metric profile was compared to 100 individuals from the CSII dataset.

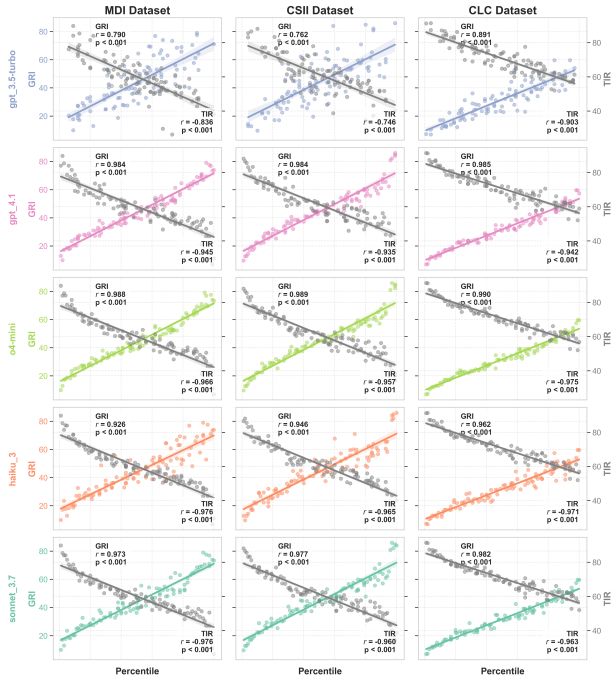

Figure 4: Scatterplots showing the relationship of the percentile rank to the GRI (y-axis left) and to TIR (y-axis right). The columns represent a distinct insulin regime datasets. Plots are labelled with Spearman correlation coefficients.

tion analysis used only the percent time spent in severe hypoglycemia and severe hyperglycemia in the regression. See Appendix E for details.

Robustness analysis. We conducted two additional analyses to assess the stability of the LLM responses. In the first analysis, we tested the effect of order reversal. We did this by swapping the order of every pairwise comparison for all of the LLM alignment experiments and reporting the consistency between the decisions and correlation of the percentile ranking with the GRI and TIR. In the second analysis, we tested the consistency of each model for three repeated runs of identical prompts for the CSII dataset. These analyses are detailed in Appendix D.

4. Results

4.1. Correlation of LLM rankings with GRI

In Figure 4, we show the results of the GRI alignment experiments for all three datasets. We show the scatterplot of the rankings for each model below. For the CLC dataset, o4-mini had the highest correlation to the GRI at 0.99 ($p < 0.001$), notably higher than the correlation to TIR (-0.975, $p < 0.001$). The results for o4-mini were similar for the MDI and CSII dataset, with o4-mini having the highest correlation to the GRI of all models.

Of the Anthropic models, Claude 3.7 Sonnet generally had higher correlation to the GRI across the three datasets than Claude 3 Haiku. Notably, Claude 3 Haiku generally had higher correlation with TIR than the GRI. These results highlight a generational shift in alignment: newer models (o4-mini, Claude 3.7 Sonnet) are closer to expert-derived indices like the GRI, whereas older models (e.g., Claude 3 Haiku) show stronger alignment to TIR.

4.2. Concordance

Figure 5 shows the concordance of pairwise rankings of the 4950 decisions for each dataset. Here, agreement with GRI indicates a model that selected the profile that with the lower GRI, and agreement with TIR indicates a model that selected the profile with the higher TIR. The agreement between GRI and TIR indicates instances where the profile with the higher TIR had the lower GRI. The profile with the higher TIR had a lower GRI 92% of the time for the CLC dataset, 90% of the time for the MDI dataset,

and 89% of the time for the CSII dataset. In all datasets, o4-mini had the highest agreement rate with the GRI at 96%, 95%, and 95% respectively. Both o4-mini and Claude 3.7 Sonnet had high agreement with TIR, but higher agreement with the GRI in all datasets. This pattern reinforces the closer alignment of newer models with expert-derived metrics such as the GRI, compared to older models that align more strongly with TIR.

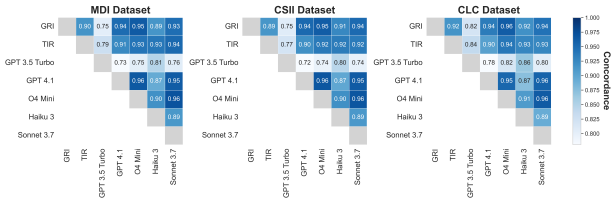

Figure 5: Concordance of models across 4950 decisions in each dataset. GRI and TIR decisions indicate instances where the metric was higher (TIR) or lower (GRI).

4.3. Case Studies

While all LLMs seem to be highly correlated with clinical metrics on average, we highlight a few instances where model preferences differed. In Figure 6, we show two examples of pairwise comparisons where at least one model disagreed with the GRI.

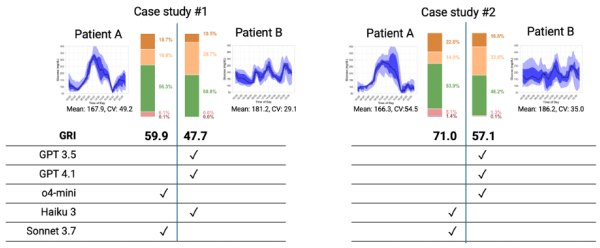

Figure 6: Case studies illustrating disagreements between LLMs and the GRI. The seven CGM-derived metrics are shown for each patient profile. The underlined GRI indicates the lower GRI, and the checkmarks indicate the "better" profile as determined by each LLM.

Case Study #1 compares a profile with a percentage of time spent in hypoglycemia of 6.2% and a percentage of TIR of 56.3% to a profile with no hypoglycemia (0%) and a percentage of TIR of 50.8%. Claude 3.7 Sonnet and o4-mini preferred the profile with higher TIR, despite the higher time in hypoglycemia, which had a higher GRI (59.9 vs. 47.7).

Case Study #2 compares Profile A, which has higher time in range (TIR) but also a high percentage of time spent in hypoglycemia, to Profile B, which has lower TIR but less hypoglycemia. Both Claude 3 Haiku and Claude 3.7 Sonnet preferred Profile A, despite its substantially higher GRI (71.0 vs. 57.1). While Profile A had a higher TIR than Profile B, it had a higher percentage of time spent in severe hypoglycemia, hypoglycemia, and severe hyperglycemia than Profile B. These examples highlight specific areas where LLM judgments differ from expectations, raising questions about how models internally weigh trade-offs between hypoglycemia and hyperglycemia.

4.4. Perturbation Analysis

Figure 7 shows the results from the perturbation analysis, where we create patient profiles by trading off hyperglycemia with hypoglycemia and assess which profiles were rated "better" than 100 patients. The results from the four models are compared to the references of TIR and GRI computed for each of the profiles in the grid. In panel A, we show the results from perturbations of mild hypoglycemia and mild hyperglycemia. Notably, the heatmaps show that Claude 3.7 Sonnet and o4-mini were not as sensitive to mild hypoglycemia as expected given the corresponding GRI. For example, when ranking the profile (0%, 0%) compared to the profile (0% mild hypoglycemia, 9% mild hyperglycemia), Claude 3.7 Sonnet's rank changes from 100 to 99, o4-mini's rank changes from 100 to 95, yet the GRI changes from 8 to 30. Visually, the contours of the heatmap also reveal the sensitivity to both mild hypoglycemia and mild hyperglycemia. We observe that the contours for Claude 3.7 Sonnet, Claude 3 Haiku, and o4-mini all resembled the contour of TIR, and not GRI. In other words, the models appear to be equally sensitive to mild hyperglycemia as to mild hypoglycemia, while the GRI penalizes mild hypoglycemia more than mild hyperglycemia. This reveals divergence between the GRI and LLMs that was not evident from previous analyses such as correlation and concordance.

On the other hand, this inconsistency with the GRI was not as pronounced in perturbations of severe hypoglycemia and severe hyperglycemia. In Panel B, Claude 3.7 Sonnet and o4-mini appeared to be more sensitive to severe hypoglycemia than severe hyper-

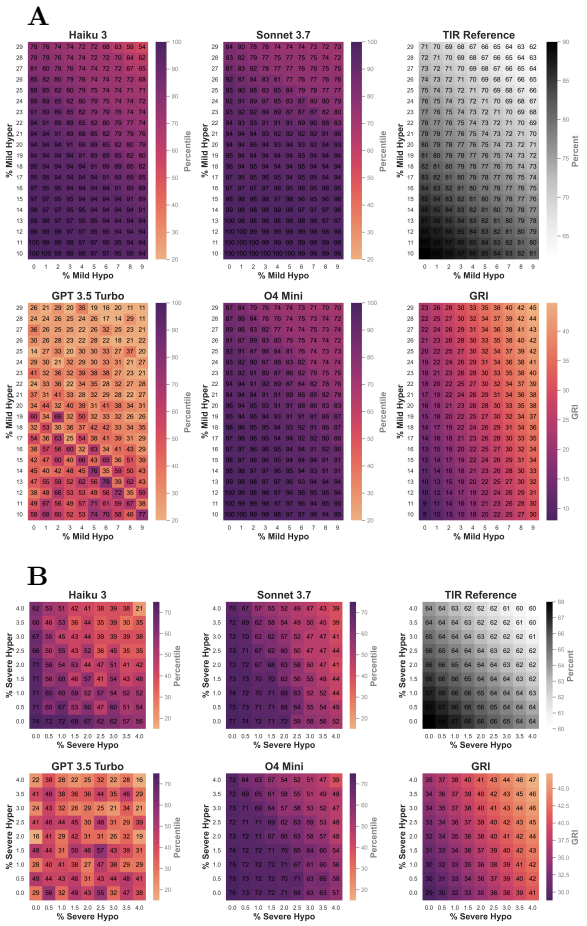

Figure 7: Perturbation analysis. (A) Mild hypoglycemia vs. mild hyperglycemia. (B) Severe hypoglycemia vs. severe hyperglycemia. For all models and GRI, we show the total number of individuals from the CSII dataset that the profile rated was "better" than compared to the CGM profile in the cell. TIR is provided as a reference.

glycemia, as indicated by the steeper descent in ranking along the x-axis than the y-axis, similar to the GRI. This trend was less obvious for Claude 3 Haiku, where the model's sensitivity to hypoglycemia over hyperglycemia was not as clearly visualized.

4.5. Regressing LLM Preference Weights

4.5.1. Regression on dataset rankings

In Figure 8, we show the results of the regression analysis where the dependent variable is the derived percentile rank, the covariates are % Severe Hypo, % Mild Hypo, % Mild Hyper, and % Severe Hyper, and observations include all 100 patients in a given dataset. Rather than focusing on the absolute coefficient values, we focus on evaluating their ratios as a measure of the relative importance.

In the GRI, the ratio of coefficients for mild hyperglycemia (0.8) and severe hyperglycemia (1.6) is 1:2, implying that each percentage of time spent in severe hyperglycemia is considered twice as bad as time spent in mild hyperglycemia. Across all three datasets, the coefficients generally have higher values for the time spent in severe hyperglycemia as compared to time spent in mild hyperglycemia. In fact, for all LLMs except for GPT 3.5, the coefficient ratio is very close to 1:2 implying strong consistency with the GRI. GPT 3.5 has a more 1:1 ratio, which is more closely reflective of TIR tradeoffs.

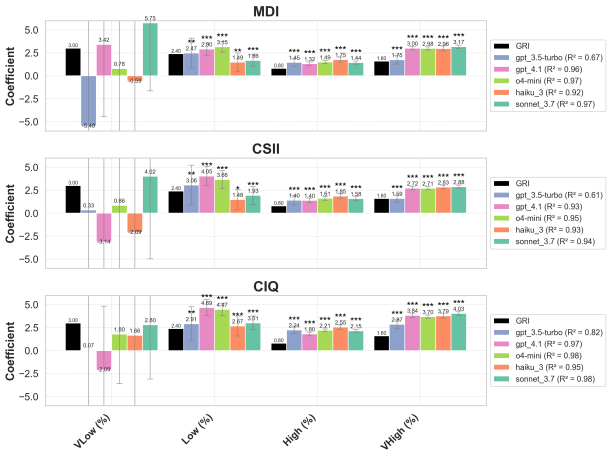

Figure 8: Regression analysis of the four out-of-range indices and the percentile ranking. * < .05; ** < .01; *** < .001

We similarly examine the relative risk of mild hypoglycemia to mild hyperglycemia. The GRI weights time spent in mild hypoglycemia three times as much as time spent in mild hyperglycemia. While GPT 4.1 and o4-mini reflect greater sensitivity to mild hypoglycemia, other models do not. The confidence intervals for the regression coefficients for time spent in severe hypoglycemia were large. This was likely due to the fact that there were many profiles in the simulated datasets with no time spent in severe hypoglycemia.

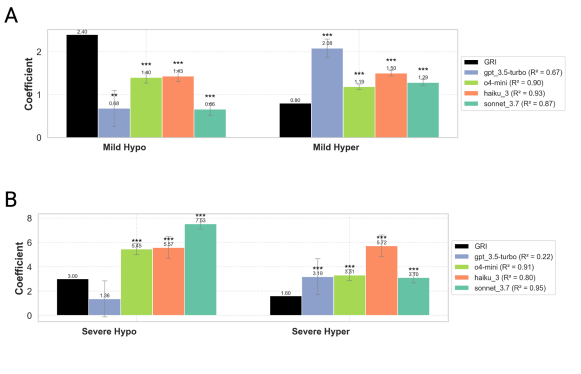

Figure 9: Regression analysis including perturbed profile. A positive coefficient indicates contribution to the perturbation score, where a higher score indicates worse glycemic control. Panels correspond to coefficients from perturbation analysis of (A) mild hypoglycemia and mild hyperglycemia and (B) severe hypoglycemia and severe hyperglycemia. The legend indicates the model and regression linear fit. * < .05; ** < .01; *** < .001

4.5.2. Regression on datasets and perturbed profiles rankings

In Figure 9, we show regression analysis where the perturbation score is regressed on the two variables that were perturbed while controlling for the remaining two variables that were held constant in the perturbation analysis. Observations include the perturbed profiles and their ranking against the 100 profiles in the CSII dataset.

In 9A, we show the coefficients trading-off between mild hypoglycemia and mild hyperglycemia. The GRI assigns thrice the weight to mild hypoglycemia as it does to mild hyperglycemia. However for all models except o4-mini, this ratio is less than one, indicating higher relative importance placed on mild hyperglycemia. In 9B, we show coefficients weighing severe hypoglycemia against severe hyperglycemia. The GRI considers severe hypoglycemia to be roughly twice the weight as severe hyperglycemia. While Claude 3.7 Sonnet and o4-mini assign the same relative importance, GPT 3.5 and Claude 3 Haiku do not.

4.6. Robustness Analysis

Our robustness analysis revealed there was an effect from the order of the pairwise comparison, but overall alignment to the GRI and TIR remained similar for most models. When evaluating the effect of rerunning the same prompts multiple times, we found all models were generally stable, with the exception of GPT 3.5. The results from both of these analyses can be found in Appendix D.

5. Discussion

We presented a novel evaluation framework designed to reveal implicit value judgments within LLMs when no gold standard answer exists. Our methodology can be generalized to any consensus-based medical decision to evaluate alignment with complex clinical judgments. We focused on the setting of assessing the quality of glycemic control from CGM data in diabetes. Using the GRI as a proxy for expert judgment, we assessed how well commercial LLMs contextualize multiple glycemic metrics. This was followed by a perturbation analysis designed to reveal the LLMs' relative weighting of hypo- and hyperglycemia.

Our findings suggest a generational shift in value alignment. Older models, GPT-3.5 and Claude 3 Haiku, were more closely aligned to TIR than to the GRI. Newer models, such as o4-mini and Claude 3.7 Sonnet, are more closely aligned to the GRI. Since the GRI was not introduced until 2023, it is possible that the older generation of LLMs were not trained on literature that used this metric when assessing glycemic control. Another plausible explanation is that newer models like o4-mini leverage advanced reasoning that enables superior clinical reasoning. Since commercial LLMs are constantly undergoing reinforcement learning from human feedback, evolving model behavior is expected and has been documented previously (Chen et al., 2024). This variation raises questions about the stability and transparency of LLM alignment over time. The fact that newer models better reflect clinical priorities could signal progress in clinical alignment. However, without transparency into proprietary alignment datasets, it is unclear whether these changes reflect deliberate improvements. This lack of transparency challenges the feasibility of long-term clinical trust.

Although LLMs demonstrated high correlation and concordance with expert-derived metrics, we identified specific instances of disagreement. To better understand how models evaluate glycemic extremes, we conducted a perturbation analysis, a form of explainable AI, involving over 100,000 LLM API calls. Notably, GPT-3.5 generated unexpected and incon-

sistent responses, in contrast to other LLMs, which appeared to provide more stable responses. While the GRI places significantly greater weight on mild hypoglycemia than mild hyperglycemia, our perturbation analysis revealed that some LLMs tended to weigh mild hyperglycemia and mild hypoglycemia equally. These inconsistencies were less pronounced in trade-offs between severe hypoglycemia and severe hyperglycemia, where models showed more sensitivity to time spent in severe hypoglycemia than severe hyperglycemia, which was in line with the GRI. Given the acuity of hypoglycemia, this behavior deviates substantially from clinical consensus and raises concerns about the clinical judgment embedded in current LLMs.

Our work has several limitations. First, the GRI is an approximation of clinician consensus. Although it correlates strongly with expert assessments, the GRI does not perfectly fit the aggregate clinician rankings it was derived from (Klonoff et al., 2023). Consequently, differences between LLM-rankings and the GRI may reflect limitations in the GRI itself rather than true misalignment of the models. Future work should include direct comparisons of LLMs with individual clinicians to better ground LLM evaluation in real-world practice. Second, our technical evaluation focused on commercial LLMs using a standardized, minimal prompt format. We did not explore open-source LLMs, sensitivity to prompt engineering, variations in phrasing, or inclusion of contextual information (e.g., treatment regimen). Nor did we ask models to explain their reasoning, which might have altered the responses and given insight into internal model reasoning. Future work should explore how LLM judgements might shift under more tailored prompting strategies. Third, we used simulated CGM data from synthetic data. This was done in order to enable controlled, large-scale evaluations of commercial LLMs without the constraints of patient privacy. Although synthetic CGM data is widely used in diabetes technology research (Cobelli and Kovatchev, 2023), replication using real CGM data is needed, particularly for edge cases that may be underrepresented in simulation environments. Together, these limitations highlight the need for continued development of evaluation frameworks that incorporate real clinician input, richer clinical context, and real-world patient data.

6. Conclusion

In conclusion, we demonstrated that even subtle inconsistencies in the alignment of LLMs for clinical decision-making can result in recommendations that conflict with safety norms. This demonstrated an important challenge in deploying AI systems in medicine: models may be misaligned with the consensus-based reasoning that clinicians rely on when making clinical judgements. We show that it is important to investigate not only the generated output of LLMs, but also the underlying value systems driving outputs. The framework we have presented in our work can be extended to other applications for the evaluation of clinical LLMs in settings where there are subjective trade-offs. Looking forward, we argue for the evolution of regulatory frameworks to include value-informed, explainable AI, where evaluations probe whether LLMs encode clinical priorities consistent with real-world expert reasoning.

Acknowledgments

This work was partially supported by T32HD040128 from the NICHD/NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Figure 1 and Figure 3 were adapted from BioRender figure https://BioRender.com/pur31q4.
Author disclosures are as follows: D.C.K. is a consultant for Afon, Atropos Health, Embecta, GlucoTrack, Lifecare, Novo, SynchNeuro, and Thirdwayv. MVT has received research support and royalties from Dexcom and Tandem Diabetes Care handled by the University of Virginia's Licensing and Ventures Group.

References

J Stuart Ablon and Enrico E Jones. Validity of controlled clinical trials of psychotherapy: findings from the NIMH treatment of depression collaborative research program. *Am. J. Psychiatry*, 159 (5):775–783, May 2002.

ADVANCE Collaborative Group, Anushka Patel, Stephen MacMahon, John Chalmers, Bruce Neal, Laurent Billot, Mark Woodward, Michel Marre, Mark Cooper, Paul Glasziou, Diederick Grobbee, Pavel Hamet, Stephen Harrap, Simon Heller, Lisheng Liu, Giuseppe Mancia, Carl Erik Mogensen, Changyu Pan, Neil Poulter, Anthony

Rodgers, Bryan Williams, Severine Bompoint, Bastiaan E de Galan, Rohina Joshi, and Florence Travert. Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. *N. Engl. J. Med.*, 358(24):2560–2572, June 2008.

Grazia Aleppo. Clinical application of time in range and other metrics. *Diabetes Spectr.*, 34(2):109–118, May 2021.

R Andrew Taylor. AI agents, automaticity, and value alignment in health care. *NEJM AI*, July 2025.

Anthropic. Anthropic claude API, 2025a.

Anthropic. Introducing the next generation of claude, 2025b. Cited July 23, 2025; Available from: https://www.anthropic.com/news/claude-3-family.

Tadej Battelino, Thomas Danne, Richard M Bergenstal, Stephanie A Amiel, Roy Beck, Torben Biester, Emanuele Bosi, Bruce A Buckingham, William T Cefalu, Kelly L Close, Claudio Cobelli, Eyal Dassau, J Hans DeVries, Kim C Donaghue, Klemen Dovc, Francis J Doyle, 3rd, Satish Garg, George Grunberger, Simon Heller, Lutz Heinemann, Irl B Hirsch, Roman Hovorka, Weiping Jia, Olga Kordonouri, Boris Kovatchev, Aaron Kowalski, Lori Laffel, Brian Levine, Alexander Mayorov, Chantal Mathieu, Helen R Murphy, Revital Nimri, Kirsten Nørgaard, Christopher G Parkin, Eric Renard, David Rodbard, Banshi Saboo, Desmond Schatz, Keaton Stoner, Tatsuiko Urakami, Stuart A Weinzimer, and Moshe Phillip. Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range. *Diabetes Care*, 42(8):1593–1603, August 2019.

L I Bonchek. Sounding board. are randomized trials appropriate for evaluating new operations? *N. Engl. J. Med.*, 301(1):44–45, July 1979.

Laura E Bothwell, Jeremy A Greene, Scott H Podolsky, and David S Jones. Assessing the gold standard–lessons from the history of RCTs. *N. Engl. J. Med.*, 374(22):2175–2181, June 2016.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv [cs.CL]*, May 2020.

Maria Ana Cardei, Josephine Lamp, Mark Derdzinski, and Karan Bhatia. DM-bench: Benchmarking LLMs for personalized decision making in diabetes management. *arXiv [cs.LG]*, October 2025.

Lingjiao Chen, Matei Zaharia, and James Zou. How is ChatGPT's behavior changing over time? *Issue 6.2, Spring 2024*, 6(2), March 2024.

Hee Jung Choi and Shriti Raj. Episode-driven insights: Can large language models tackle multimodal diabetes data? *medRxiv*, page 2025.04.24.25326385, April 2025.

Claudio Cobelli and Boris Kovatchev. Developing the UVA/padova type 1 diabetes simulator: Modeling, validation, refinements, and utility. *J. Diabetes Sci. Technol.*, 17(6):1493–1505, November 2023.

Benjamin Djulbegovic and Gordon Guyatt. Evidence vs consensus in clinical practice guidelines. *JAMA*, 322(8):725–726, August 2019.

Klemen Dovc and Tadej Battelino. Time in range centered diabetes care. *Clin. Pediatr. Endocrinol.*, 30(1):1–10, January 2021.

Nuha A ElSayed, Grazia Aleppo, Vanita R Aroda, Raveendhara R Bannuru, Florence M Brown, Dennis Bruemmer, Billy S Collins, Marisa E Hilliard, Diana Isaacs, Eric L Johnson, Scott Kahan, Kamlesh Khunti, Jose Leon, Sarah K Lyons, Mary Lou Perry, Priya Prahalad, Richard E Pratley, Jane Jeffrie Seley, Robert C Stanton, and Robert A Gabbay. 6. glycemic targets: Standards of care in Diabetes—2023. *Diabetes Care*, 46(Supplement_1): S97–S110, December 2022.

Thomas R Frieden. Evidence for health decision making - beyond randomized, controlled trials. *N. Engl. J. Med.*, 377(5):465–475, August 2017.

Elizabeth Healey, Amelia L M Tan, Kristen L Flint, Jessica Ruiz, and Isaac S Kohane. A case study on using a large language model to analyze continuous glucose monitoring data. *Sci. Rep.*, 15(1): 1143, January 2025.

International Diabetes Center. Ambulatory glucose profile: AGP reports. Availablefrom:http://www.agpreport.org/agp/agpreports.

David C Klonoff, Jing Wang, David Rodbard, Michael A Kohn, Chengdong Li, Dorian Liepmann, David Kerr, David Ahn, Anne L Peters, Guillermo E Umpierrez, Jane Jeffrie Seley, Nicole Y Xu, Kevin T Nguyen, Gregg Simonson, Michael S D Agus, Mohammed E Al-Sofiani, Gustavo Armaiz-Pena, Timothy S Bailey, Ananda Basu, Tadej Battelino, Sewagegn Yeshiwas Bekele, Pierre-Yves Benhamou, B Wayne Bequette, Thomas Blevins, Marc D Breton, Jessica R Castle, James Geoffrey Chase, Kong Y Chen, Pratik Choudhary, Mark A Clements, Kelly L Close, Curtiss B Cook, Thomas Danne, Francis J Doyle, 3rd, Angela Drincic, Kathleen M Dungan, Steven V Edelman, Niels Ejskjaer, Juan C Espinoza, G Alexander Fleming, Gregory P Forlenza, Guido Freckmann, Rodolfo J Galindo, Ana Maria Gomez, Hanna A Gutow, Lutz Heinemann, Irl B Hirsch, Thanh D Hoang, Roman Hovorka, Johan H Jendle, Linong Ji, Shashank R Joshi, Michael Joubert, Suneil K Koliwad, Rayhan A Lal, M Cecilia Lansang, Wei-An Andy Lee, Lalantha Leelarathna, Lawrence A Leiter, Marcus Lind, Michelle L Litchman, Julia K Mader, Katherine M Mahoney, Boris Mankovsky, Umesh Masharani, Nestoras N Mathioudakis, Alexander Mayorov, Jordan Messler, Joshua D Miller, Viswanathan Mohan, James H Nichols, Kirsten Nørgaard, David N O'Neal, Francisco J Pasquel, Athena Philis-Tsimikas, Thomas Pieber, Moshe Phillip, William H Polonsky, Rodica Pop-Busui, Gerry Rayman, Eun-Jung Rhee, Steven J Russell, Viral N Shah, Jennifer L Sherr, Koji Sode, Elias K Spanakis, Deborah J Wake, Kayo Waki, Amisha Wallia, Melissa E Weinberg, Howard Wolpert, Eugene E Wright, Mihail Zilbermint, and Boris Kovatchev. A glycemia risk index (GRI) of hypoglycemia and hyperglycemia for continuous glucose monitoring validated by clinician ratings. *J. Diabetes Sci. Technol.*, 17(5):1226–1242, September 2023.

Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA type 1 diabetes simulator: New features. *J. Diabetes Sci. Technol.*, 8(1): 26–34, January 2014.

Anthony L McCall. Insulin therapy and hypoglycemia. *Endocrinol. Metab. Clin. North Am.*, 41 (1):57–87, March 2012.

Revital Nimri, Eyal Dassau, Tomer Segall, Ido Muller, Natasa Bratina, Olga Kordonouri, Rachel Bello, Torben Biester, Klemen Dovc, Ariel Tenenbaum, Avivit Brener, Marko Šimunović, Sophia D Sakka, Michal Nevo Shenker, Caroline Gb Passone, Irene Rutigliano, Davide Tinti, Clara Bonura, Silvana Caiulo, Anna Ruszala, Barbara Piccini, Dinesh Giri, Ronnie Stein, Ivana Rabbone, Patrizia Bruzzi, Jasna Šuput Omladič, Caroline Steele, Guglielmo Beccuti, Michal Yackobovitch-Gavan, Tadej Battelino, Thomas Danne, Eran Atlas, and Moshe Phillip. Adjusting insulin doses in patients with type 1 diabetes who use insulin pump and continuous glucose monitoring: Variations among countries and physicians. *Diabetes Obes. Metab.*, 20(10):2458–2466, October 2018.

OpenAI. OpenAI API: v1/responses endpoint, 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey,

Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C J Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 technical report. *arXiv [cs.CL]*, March 2023.

Juliessa M Pavon, David Schlientz, Matthew L Maciejewski, Nicoleta Economou-Zavlanos, and Richard H Lee. Large language models in diabetes management: The need for human and artificial intelligence collaboration. *Diabetes Care*, 48(2):182–184, February 2025.

Inioluwa Deborah Raji, Roxana Daneshjou, and Emily Alsentzer. It's time to bench the medical exam benchmark. *NEJM AI*, 2(2), January 2025.

David Rodbard. Quality of glycemic control: Assessment using relationships between metrics for safety and efficacy. *Diabetes Technol. Ther.*, 23(10):692–704, October 2021.

Nicole L Spartano, Brenton Prescott, Maura E Walker, Eleanor Shi, Guhan Venkatesan, David Fei, Honghuang Lin, Joanne M Murabito, David Ahn, Tadej Battelino, Steven V Edelman, G Alexander Fleming, Guido Freckmann, Rodolfo J Galindo, Michael Joubert, M Cecilia Lansang, Julia K Mader, Boris Mankovsky, Nestoras N Mathioudakis, Viswanathan Mohan, Anne L Peters, Viral N Shah, Elias K Spanakis, Kayo Waki, Eugene E Wright, Mihail Zilbermint, Howard A Wolpert, and Devin W Steenkamp. Expert clinical interpretation of continuous glucose monitor reports from individuals without diabetes. *J. Diabetes Sci. Technol.*, page 19322968251315171, February 2025.

SPRINT Research Group, Jackson T Wright, Jr, Jeff D Williamson, Paul K Whelton, Joni K Snyder, Kaycee M Sink, Michael V Rocco, David M Reboussin, Mahboob Rahman, Suzanne Oparil, Cora E Lewis, Paul L Kimmel, Karen C Johnson, David C Goff, Jr, Lawrence J Fine, Jeffrey A Cutler, William C Cushman, Alfred K Cheung, and Walter T Ambrosius. A randomized trial of in-

tensive versus standard blood-pressure control. *N. Engl. J. Med.*, 373(22):2103–2116, November 2015.

The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N. Engl. J. Med.*, 329(14):977–986, September 1993.

Roberto Visentin, Enrique Campos-Náñez, Michele Schiavon, Dayu Lv, Martina Vettoretti, Marc Breton, Boris P Kovatchev, Chiara Dalla Man, and Claudio Cobelli. The UVA/padova type 1 diabetes simulator goes from single meal to single day. *J. Diabetes Sci. Technol.*, 12(2):273–281, March 2018.

Kun-Hsing Yu, Elizabeth Healey, Tze-Yun Leong, Isaac S Kohane, and Arjun K Manrai. Medical artificial intelligence and human values. *N. Engl. J. Med.*, 390(20):1895–1904, May 2024.

Appendix A. Data

We simulated the following three insulin modalities:

1. **MDI**: Multiple daily injections regimen simulated with a daily long-acting insulin basal dose and fast-acting insulin boluses provided at mealtimes.

2. **CSII**: Continuous subcutaneous insulin infusion (open-loop pump therapy) simulated using individualized therapy profiles.

3. **CLC**: Closed-loop control therapy simulated using the USS-Virginia algorithm, the academic version of a commercially available automated insulin delivery system (Control-IQ).

We simulated patient data using the UVA/Padova T1D FDA-accepted patient simulator (Visentin et al., 2018). For each dataset, we generated 14 consecutive days of CGM data for each of the 100 synthetic adults. Each day included three meals and three snacks. To simulate realistic conditions, both behavioral and metabolic variability were introduced. Meal timing varied within typical breakfast (7:00), lunch (13:00), and dinner (19:00) windows (±1.5 h), with snacks scheduled between meals (±30 min). Carbohydrate amounts for meals and snacks were sampled from ranges proportional to body weight. Additional behavioral variability included carbohydrate counting errors of up to 20% and random meal announcements delays ranging from -15 to +15 minutes. There was also induced intra-day and inter-day metabolic variability represented by fluctuations in insulin sensitivity related parameters. Descriptive statistics and distributions for each dataset are shown in Table 1 and Figure 10, respectively.

Table 1: Statistics of CGM data by dataset. Mean and standard deviation are shown.

	MDI	CSII	CLC
Number of Patients	100	100	100
Length of CGM data (days)	14 ± 0	14 ± 0	14 ± 0
GRI	43.95 ± 16.38	44.37 ± 16.75	31.86 ± 12.78
Mean (mg/dL)	165.51 ± 15.04	165.82 ± 16.81	154.12 ± 8.78
CV (%)	34.71 ± 8.24	36.65 ± 6.95	33.38 ± 7.31
TIR (%)	59.95 ± 11.24	61.17 ± 11.19	70.60 ± 8.84
VLow (%)	0.08 ± 0.19	0.09 ± 0.21	0.09 ± 0.20
Low (%)	2.46 ± 2.45	2.45 ± 1.94	1.97 ± 1.68
High (%)	27.76 ± 7.68	24.81 ± 6.56	21.10 ± 5.45
VHigh (%)	9.75 ± 7.65	11.48 ± 8.59	6.24 ± 5.48

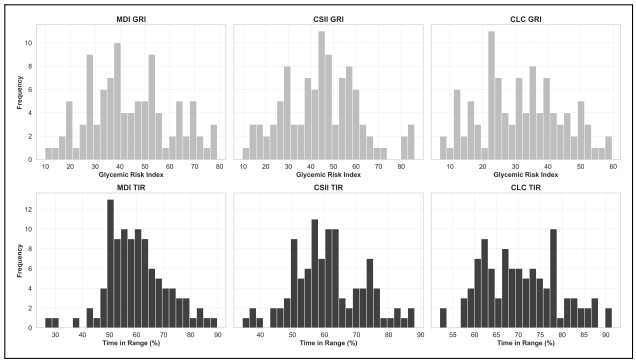

Figure 10: GRI distributions by dataset.

Appendix B. Measuring LLM Alignment with the Glycemic Risk Index

We used three OpenAI GPT series models (OpenAI et al., 2023; Brown et al., 2020) and two Anthropic Claude series models (Anthropic, 2025b). Our evaluation used five commercial LLMs from two companies Table 2).

Table 2: Description of LLMs models evaluated.

Abbrev.	Full Model Name	Release Date	Training Cutoff
sonnet3.7	claude-3-7-sonnet-20250219	2025-02-19	Nov 2024
haiku3	claude-3-haiku-20240307	2024-03-07	Aug 2023
gpt3.5-turbo	gpt-3.5-turbo-0125	2023-01-25	Aug 2021
gpt4.1	gpt-4.1-2025-04-14	2025-04-14	May 2024
o4-mini	o4-mini-2025-04-16	2025-04-16	May 2024

For each patient, seven metrics were extracted from the CGM data. These metrics were then used for pairwise comparisons to all other patients in the dataset. The prompt used for the comparisons is shown in Figure 2. Each dataset had 100 patients, corresponding to 4950 total comparisons. The LLM was prompted to identify which patient had better glycemic control. Each patient was then assigned a score based on the number of other patients the LLM deemed them "better than", yielding a score from 0 to 99. Rankings were then assembled by subtracting this score from 100.

Appendix C. Perturbation Analysis

To better understand how the models place value on hypoglycemia and hyperglycemia, we performed a perturbation analysis. In these experiments, we perturbed the metrics by reallocating percentages from

one metric to another and making pairwise comparisons to see which profiles the models deemed "better." We leveraged the same prompt structure from Figure 2, except we removed Mean and CV from the prompt and only included the time in glycemic ranges. We created artificial profiles across a range of hypoglycemia and hyperglycemia levels. The goal of the perturbation analysis was to elucidate the values embedded in the LLM for hypoglycemia and hyperglycemia.

We did this by prompting the LLM to perform pairwise comparisons of metric profiles with perturbations on hypoglycemia and hyperglycemia. Figure 7 shows an overview of the experimental setup. By asking the models to decide which profile was "better," we aimed to reveal how the models weigh hypoglycemia and hyperglycemia when assessing quality of glycemic control. We used the 100 individuals from the CSII dataset to compare each profile to. Each profile was assessed based on how many of the 100 individuals the model deemed it was better than.

We performed two perturbation experiments:

1. **Mild hypoglycemia vs. mild hyperglycemia:** Mild hypoglycemia was varied from 0% to 9% and mild hyperglycemia from 10% to 29%, both in 1% increments. Severe hyperglycemia and severe hypoglycemia were held constant at 0%. This resulted in 200 profiles, corresponding to 20,000 API requests per model.

2. **Severe hypoglycemia vs. severe hyperglycemia:** Severe hypoglycemia was varied from 0% to 4% and severe hyperglycemia from 0% to 4%, both in 0.5% increments. Mild hypoglycemia was held constant at 2% and mild hyperglycemia at 30%.

Due to the computational cost of each perturbation experiment, we reduced the number of LLMs studied by evaluating two older models (GPT-3.5 and Claude 3 Haiku) and two newer models (o4-mini and Claude 3.7 Sonnet).

Appendix D. Robustness Analysis

D.0.1. Methodology

Testing the effect of order reversal. To better understand the effect of the order of the comparisons in the prompts, we conducted an additional analysis where we reran the alignment experiments with every pairwise comparison appearing in the reversed order. For example, if in the original experiments patient X was first in the prompt when patient X was compared to patient Y, the reversed prompt put patient Y first.

Testing the stability of repeated runs. We further evaluated the stability of answers in repeated runs using the same prompts. In this analysis, we focused on the CSII dataset and ran the original experiment three additional times. For the OpenAI models, these were ran three additional times in the reversed order. For the Anthropic models, these were ran three additional times in the non-reversed order.

D.0.2. Results

Results of order reversal

In Table 3, we show the alignment of the models when each pairwise comparison is reversed. The concordance represents the overall agreement of each of the 4950 pairs between the original prompts and the prompts with the order reversed.

Results of multiple iterations In Table 4, we show the results from three additional repeated runs for the CSII dataset.

Appendix E. Statistical Analysis and Software

Language models. OpenAI models were leveraged using the `v1/responses` endpoint in batch mode (Version 1.84.0) (OpenAI, 2025). Claude models were leveraged using Anthropic's Message Batch API (Version 0.55.0) (Anthropic, 2025a).

Comparison to GRI. We computed the GRI for each patient. For each dataset, we measured the correlation of the LLM-ranked percentiles and the GRI using the Spearman correlation coefficient. We also measured the correlation with the other seven metrics, including TIR. This analysis was conducted using `scipy` (Version 1.13.0).

Regression analysis. Regression analysis was conducted for each model using `statsmodels` (Version 0.14.2), allowing for a y-intercept, and plotting confidence intervals for each covariate. All analyses were performed in Python (Version 3.12.3). The regression for all analyses used the ranking as the output, where a higher rank corresponded to a worse profile. This was computed by counting the total number of profiles that the profile was seen as "better" than and subtracting that number from 100.

Table 3: Results from order reversal. Table shows the correlation of the LLM percentile rank with the GRI and TIR for each model and dataset. The Reversed column indicates whether the pairwise comparison was reversed. The concordance represents the percent agreement between the reversed results and the non-reversed result for each model and dataset.

Dataset	Model	Reversed	GRI corr	TIR corr	Concordance
MDI	haiku_3	No	0.93	-0.98	0.84
		Yes	0.90	-0.97	
	sonnet_3.7	No	0.97	-0.98	0.94
		Yes	0.97	-0.98	
	gpt_3.5-turbo	No	0.79	-0.84	0.60
		Yes	0.76	-0.81	
	gpt_4.1	No	0.98	-0.95	0.89
		Yes	0.99	-0.95	
	o4-mini	No	0.99	-0.97	0.95
		Yes	0.99	-0.96	
CSII	haiku_3	No	0.95	-0.96	0.84
		Yes	0.93	-0.96	
	sonnet_3.7	No	0.98	-0.96	0.93
		Yes	0.98	-0.96	
	gpt_3.5-turbo	No	0.76	-0.75	0.56
		Yes	0.72	-0.76	
	gpt_4.1	No	0.98	-0.93	0.88
		Yes	0.99	-0.93	
	o4-mini	No	0.99	-0.96	0.94
		Yes	0.99	-0.96	
CIQ	haiku_3	No	0.96	-0.97	0.84
		Yes	0.93	-0.95	
	sonnet_3.7	No	0.98	-0.96	0.93
		Yes	0.99	-0.97	
	gpt_3.5-turbo	No	0.89	-0.90	0.67
		Yes	0.83	-0.87	
	gpt_4.1	No	0.98	-0.94	0.89
		Yes	0.99	-0.95	
	o4-mini	No	0.99	-0.97	0.95
		Yes	0.99	-0.98	

Table 4: Results of multiple additional runs. The results show the correlation of the LLM percentile rankings to the GRI and TIR for each additional run. All prompts were kept exactly the same for each additional run.

Model	Dataset	run	GRI corr	TIR corr
haiku_3	CSII	0	0.946	-0.962
haiku_3	CSII	1	0.945	-0.963
haiku_3	CSII	2	0.944	-0.963
sonnet_3.7	CSII	0	0.977	-0.959
sonnet_3.7	CSII	1	0.976	-0.961
sonnet_3.7	CSII	2	0.978	-0.96
gpt_3.5-turbo	CSII	0	0.68	-0.73
gpt_3.5-turbo	CSII	1	0.721	-0.763
gpt_3.5-turbo	CSII	2	0.731	-0.759
gpt_4.1	CSII	0	0.986	-0.928
gpt_4.1	CSII	1	0.986	-0.929
gpt_4.1	CSII	2	0.986	-0.928
o4-mini	CSII	0	0.99	-0.956
o4-mini	CSII	1	0.989	-0.958
o4-mini	CSII	2	0.989	-0.959