



AI-Driven Glycemic Management Using LLM Guidance and Reinforcement Learning

Internship Report

Taha RAMDAN

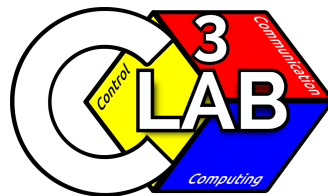
Industrial Supervisor: **Saverio MASCOLO**

PhD Collaborator and mentor: **Giada LOPS**

Polytech Supervisor: **Peter BANTON**



**Politecnico
di Bari**



Date: **June 16, 2025**

Keywords: Automated Insulin Delivery; reinforcement learning; large language models; safe control; interpretability; SimGlucose; De Paola model; time-in-range.

Acknowledgments

It is a privilege to express my profound gratitude to **Prof. Saverio Mascolo**, who welcomed me to **C3Lab** and granted me the opportunity to join his group. I am deeply thankful for his supervision, and trust throughout this internship; his example of scientific rigor and leadership has been truly inspiring.

I am especially indebted to **Giada Lops**, a PhD student at C3Lab, whose doctoral project I had the honor to join as an intern. Before my arrival in Italy, she proactively shared a curated set of key papers that provided the right foundations; during the internship, she designed and organized the research plan, guided the experimental methodology, and mentored me day-to-day on both technical and clinical aspects. Her PhD research provided the scientific backbone of this report, and her generosity and high standards shaped every stage of my work.

I would also like to thank **Prof. Peter Banton**, my Polytech supervisor, for his attentive academic follow-up and practical advice regarding the program requirements.

C3Lab has been an exceptionally welcoming place to learn and grow. I am grateful to my colleagues and for their kindness, openness to questions, and the many generous conversations that helped me clarify ideas, challenge assumptions, and improve the quality and reproducibility of this work.

Finally, I extend my heartfelt thanks to my family and friends especially for their constant support and encouragement.

Abstract

Automated insulin delivery (AID) must jointly optimize performance, safety, and interpretability. This internship report develops and evaluates knowledge-guided control for diabetes across six chapters. **Chapter 1** standardizes in silico experimentation by wrapping the UVA/Padova *SimGlucose* model in a Gymnasium-compatible interface, defining a clinically anchored reward (target 120 mg/dL with strong penalties < 70 and > 180 mg/dL), and benchmarking PPO against discretized Q-learning. **Chapter 2** introduces an LLM commentary layer that translates raw observations into guideline-aware explanations aligned with ADA glycemic targets. **Chapter 3** strengthens reasoning and safety by prompting the LLM with structured state (CGM, IOB, meals, time-of-day) and explicit dosing rules (ICR, ISF/CF, AIT), and by enforcing a JSON schema plus a two-stage semantic parser and safety checks (e.g., suspend < 70 mg/dL).

Chapter 4 demonstrates an end-to-end LLM-controlled advisor integrated with SimGlucose and PID feedback, achieving high time-in-range (e.g., $\sim 98.8\%$ TIR in a 24 h scenario) without Level-2 hypoglycemia. **Chapter 5** proposes a hybrid RL + LLM controller with a safety layer and PLGS-like logic; stress tests reveal the safety–performance trade-off (e.g., $\sim 68.6\%$ TIR with elevated hypoglycemia), motivating uncertainty-aware gating, stronger guardrails, and multi-patient validation. **Chapter 6** extends the framework to non-pharmacological control: a Gym-compatible environment based on the DePaola activity–glucose model shows that an hourly-sampled TD3 agent can maintain normoglycemia in multi-day simulations, while highlighting the need for realistic effort budgets.

Across chapters, the contribution is a reproducible stack (wrappers, prompts/rules, schema and verifier) that couples explicit clinical knowledge with learning-based control to improve transparency and safety. Limitations include single-patient scenarios in some studies and the absence of clinical trials; future work targets uncertainty-aware fusion, energy/effort costs for activity, and evaluation across cohorts, seeds, and scenarios. Taken together, the results indicate that hybrid, knowledge-guided control improves safety and interpretability in AID while providing a practical basis for reproducible in silico evaluation.

Contents

Acknowledgments	1
Abstract	1
1 Development of an AI-Driven Glycemic Control System Using Reinforcement Learning	3
1.1 Environment Setup & Dependencies	3
1.2 Scenario Configuration (Meal Planning)	3
1.3 Custom Environment Wrapper	4
1.4 Custom Reward Function	5
1.5 PPO Training	5
1.6 Evaluation	6
1.7 Visualization	6
2 LLM Recommendation from Environment Observation	8
2.1 Objective of This Step	8
2.2 LLM-in-the-Loop Methodology	8
2.3 Illustrative Output	9
2.4 Limitations and Future Work	9
3 Enhancing LLM Reasoning with Structured Patient State and Clinical Rules	10
3.1 Integrating Structured Patient State and Clinical Rules into an LLM Advisor	10
3.2 Discussion	12
4 LLM-Controlled Insulin Advisor in SimGlucose	13
4.1 Introduction & Motivation	13
4.2 Dataset Construction	13
4.2.1 Parsing Ohio T1DM XML Files	13
4.2.2 Prompt-Completion Formatting	14
4.3 Fine-Tuning Method (LoRA)	14
4.4 Semantic Parsing of LLM Output	15
4.5 Hybrid LLM + PID Controller	15
4.5.1 Algorithm Overview	15
4.6 Simulation Setup	15
4.6.1 Virtual Patient and Meal Scenario	15
4.7 Results	16

4.7.1	Glucose Profile and Controller Actions	16
4.7.2	Narrative Results	17
4.7.3	Quantitative Metrics	17
4.7.4	Discussion	18
4.7.5	Implications for Chapter 5	18
5	Hybrid Decision Making: RL Policy + LLM	19
5.1	Background and Motivation	19
5.2	Hybrid Controller (High-Level Design)	19
5.3	Experimental Protocol	19
5.3.1	Simulator and Scenario	19
5.3.2	Endpoints and Metrics	20
5.4	Results	20
5.4.1	Qualitative Traces	20
5.4.2	Quantitative Metrics (24 h, <code>adolescent#001</code>)	21
5.5	Discussion	21
5.6	Limitations and Future Work	22
6	Custom Diabetes Environment (De Paola Model)	23
6.1	Background and State of the Art	23
6.1.1	Classical Glucose–Insulin Simulators	23
6.2	Reduced De Paola Model for Activity Control	23
6.3	Methodology	24
6.3.1	Environment Implementation	24
6.3.2	Reward Function	24
6.3.3	Reinforcement-Learning Setup	24
6.4	Results	25
6.5	Discussion	25
6.5.1	Key Findings	25
6.5.2	Limitations	26
6.5.3	Conclusion	26

Chapter 1

Development of an AI-Driven Glycemic Control System Using Reinforcement Learning

1.1 Environment Setup & Dependencies

Objective. Install and configure all libraries required to run the T1D simulation and to train the RL agent, ensuring reproducibility and stability.

Dependencies. We rely on `simglucose`, an open-source Python package implementing the UVa/PADOVA physiological model for Type 1 diabetes that provides 30 virtual patients (children, adolescents, adults), emits CGM observations every three minutes, and adheres to the Gymnasium API [1, 18, 28, 19]. We also use `gymnasium` (the successor to OpenAI Gym) as our RL environment interface, `stable-baselines3` for PPO and other state-of-the-art algorithms, and the `matplotlib`, `numpy`, and `pandas` libraries for data analysis and visualization [45, 25].

1.2 Scenario Configuration (Meal Planning)

Scenario Configuration (Meal Planning)

To faithfully reproduce the principal drivers of post-prandial glucose excursions in silico, we define a three-meal regimen reflecting typical daily eating patterns. By specifying both the timing and carbohydrate load of each meal, we ensure that the simulated patient experiences realistic rises and falls in blood glucose, providing a challenging yet clinically meaningful testbed for reinforcement-learning controllers [21].

Clinical Relevance.

- **Breakfast (06:00, 30 g):** Morning meals typically provoke a pronounced dawn phenomenon in T1D patients. A 30 g carbohydrate bolus at 6 AM captures the early-day hepatic glucose output and reduced insulin sensitivity [22, 23, 24].
- **Lunch (12:00, 60 g):** Midday meals are often the largest in carbohydrate content. Sixty grams at noon models typical Western dietary habits and generates substantial post-prandial peaks.

-
- **Dinner (18:00, 40 g):** Evening intake of 40 g reflects lighter but still significant carbohydrate consumption. Late-day insulin requirements differ from daytime, testing the controller’s ability to adapt to circadian variations in insulin sensitivity [24].

Together, these three strategically spaced meals produce inter-meal and post-prandial glucose dynamics—hypoglycemic dips and hyperglycemic spikes—closely mirroring real-world Type 1 diabetes glycemic profiles. This realistic challenge is essential to evaluate and refine RL policies under conditions that a patient would actually face.

1.3 Custom Environment Wrapper

To bridge the gap between the native SimGlucose API and the Stable-Baselines3/Gymnasium ecosystem, we introduce a tailored wrapper class, `CustomT1DWrapper`. This wrapper performs three critical adaptations:

Action-space expansion. In the default SimGlucose Gym interface, the action is a *scalar basal insulin rate* (U/min). Our wrapper keeps this basal command and additionally introduces a *bolus* channel for prandial dosing. To handle realistic meal-induced excursions, we cap basal within physiologic limits and allow a bolus up to [0, 10] U per 3-minute decision step, enforced by safety clamps.

Observation-space specification. Real-world CGM readings can span from extreme hypoglycemia to severe hyperglycemia. We therefore set the observation box to [0, 1000] mg/dL, assuring full coverage of clinically relevant glycemic values.

API compatibility and custom reward. SimGlucose’s `reset()` and `step()` methods return tuples of (obs) and (obs, raw_reward, done, info) respectively. Our wrapper remaps these to the Gymnasium signature `reset() → (obs, info)` and `step() → (obs, reward, terminated, truncated, info)` [19, 20], and applies a bespoke reward function that penalizes excursions outside the target range [15, 5].

In addition, the factory function `make_custom_env`:

1. Instantiates `T1DSimEnv` with a chosen virtual patient.
2. Injects any user-defined meal scenario after initialization.
3. Wraps the environment in `CustomT1DWrapper`.
4. Vectorizes it across four parallel environments (`n_envs=4`) for efficient policy training [27, 45].

This design ensures both clinical fidelity—by preserving the UVa/PADOVA physiological model—and algorithmic performance when using off-the-shelf RL libraries [18].

1.4 Custom Reward Function

The custom reward function was crafted to balance tight glycemic control with patient safety. In particular, we set a clinical target of 120 mg/dL—commonly recommended for people with T1D—and structured the penalties as follows. Severe hypoglycemia (blood glucose below 70 mg/dL) incurs a large, safety-critical penalty to discourage the agent from driving glucose too low. Conversely, hyperglycemia above 180 mg/dL carries a moderate penalty, reflecting its contribution to long-term complications. To promote smooth and precise regulation, we then wrap these bounds in a quadratic “shaping” term around the 120 mg/dL setpoint: deviations are squared, so larger excursions (either hypo- or hyperglycemic) are disproportionately penalized, while small fluctuations near the target only incur mild costs. This design incentivizes the policy to keep glucose tightly within the safe range, yet avoids overly aggressive or oscillatory dosing [5, 15].

1.5 PPO Training

Technical Remark. Initially, we experimented with a classical Q-learning approach by discretizing both the blood-glucose state space and the insulin-dose action space. However, this discretization introduced quantization errors and limited the granularity of the controller’s decisions. After further literature review and preliminary experiments, we transitioned to Proximal Policy Optimization (PPO). PPO natively handles continuous state and action spaces, preserving the full precision of glucose readings and insulin delivery rates. Moreover, PPO’s clipped surrogate objective yields more stable and sample-efficient learning in complex, non-stationary environments such as glucose regulation—making it a superior choice compared to traditional Q-learning for our application [25, 45].

Criteria	Q-Learning	PPO (Proximal Policy Optimization)
Action/State Space	Requires discretization	Handles continuous spaces
Precision	Limited due to discretization	High precision
Stability of Learning	Less stable in complex problems	More stable and sample-efficient
Suitability for our Project	Not ideal for precise insulin control	Well-suited for continuous insulin dosing
Scalability	Poor for high-dimensional spaces	Better scalability

Figure 1.1: Comparison of Q-learning (discretized) vs. PPO (continuous) performance on early training episodes.

Configuration: The PPO agent was trained using the `MlpPolicy` multilayer perceptron with its default architecture. The learning rate was set to 3×10^{-4} to balance convergence speed and stability. Training was run for 60,000 timesteps, which—at a 3-minute control interval and 4 parallel environments—corresponds to approximately 31.25 simulated days per environment ($60,000/(480 \times 4)$). By specifying `device=auto`, the implementation automatically leverages a GPU when available and otherwise falls back to the CPU [27, 45].

1.6 Evaluation

The trained agent was evaluated over 10 independent episodes using a deterministic policy—selecting the mean action at each step to minimize stochastic noise. We report the mean cumulative reward across these episodes as the primary performance metric, and include the standard deviation of the rewards to quantify the consistency of the policy’s performance.

1.7 Visualization

Figure 1.2 shows the blood glucose trajectory over a 24-hour period under the learned PPO policy. The solid blue line depicts the CGM readings at 3-minute intervals, the green line marks the clinical target of 120 mg/dL, and the red dashed lines indicate the hypoglycemia (70 mg/dL) and hyperglycemia (180 mg/dL) danger thresholds [5, 15]. We observe that, following each meal-induced spike, the controller successfully brings glucose back toward target with only brief excursions outside the safe range.

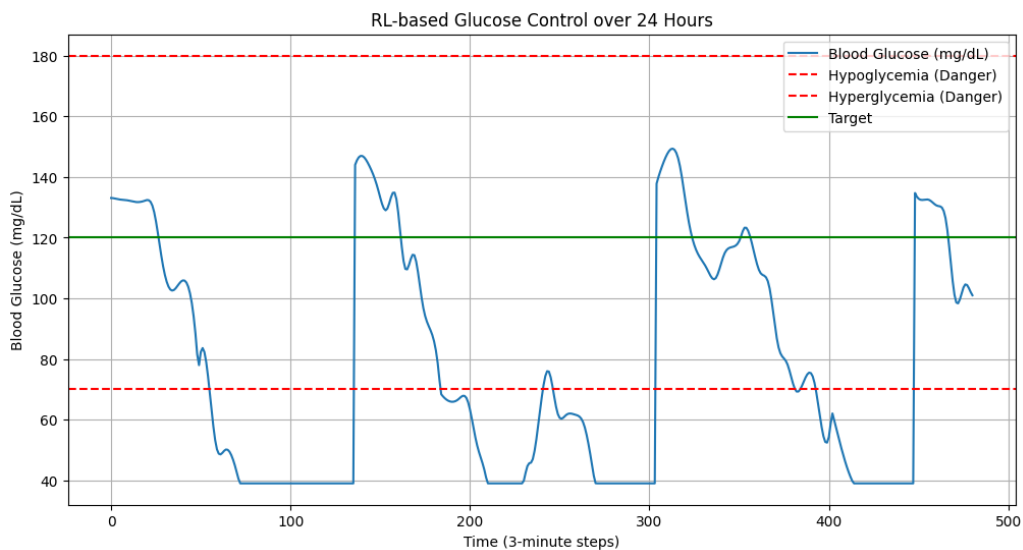


Figure 1.2: RL-based glycemic control over 24 hours.

Figure 1.3 plots the corresponding insulin infusion rates recommended by the agent. The purple trace reflects smooth, graded boluses that rise in response to postprandial glucose elevations and taper as glucose returns toward target. No abrupt or oscillatory dosing is evident, indicating a stable policy.

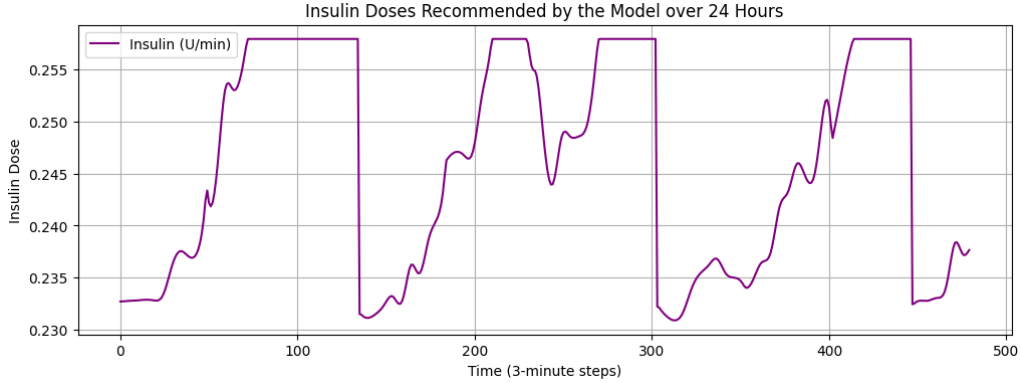


Figure 1.3: Insulin doses recommended by the model over 24 hours.

These visualizations demonstrate that the PPO agent learned to anticipate and mitigate meal-related glucose excursions with smooth, physiologically plausible insulin dosing. The policy maintains glucose predominantly within the 70–180 mg/dL safe range and converges toward the 120 mg/dL setpoint, validating the efficacy of our reinforcement-learning framework for glycemic management.

Chapter 2

LLM Recommendation from Environment Observation

2.1 Objective of This Step

The aim of this proof-of-concept is to show that a large-language model (LLM) can translate raw **CGM** observations into *interpretable*, guideline-aware advice for clinicians. Closed-loop prototypes typically output only a numeric control signal whose rationale is opaque. By attaching an LLM commentary that explicitly references recognised glycaemic targets—e.g., the American Diabetes Association recommendation of 80–130 mg/dL pre-prandial and < 180 mg/dL post-prandial [5]—each suggested action is placed in a clear clinical context, which *may improve perceived transparency* for clinicians. Textual outputs are non-prescriptive and require human oversight.

2.2 LLM-in-the-Loop Methodology

Observation–prompt mapping

At decision time t the simulator emits an observation

$$o_t = (G_t, I_t, \dot{G}_t, \dots),$$

where G_t is the **CGM glucose reading** (mg/dL) [1, 18], I_t the **basal insulin rate** (U/min) [1], and \dot{G}_t an optional rate-of-change term.¹ The observation is converted to an English prompt

Prompt(o_t) = Patient glucose $\langle G_t \rangle$ mg/dL, basal $\langle I_t \rangle$ U/min. Recommendation?,

which is fed to an *instruction-tuned* LLM (**Mistral-7B-Instruct** [6]). To ensure deterministic, reproducible outputs, inference settings are fixed to `temperature=0`, `top_p=1.0`, and `max_new_tokens=32`.

¹In the SimGlucose/UVA testbeds, the observation stream queries the CGM sensor and the default control action is a scalar basal command [1, 18].

Textual recommendation

The model returns a free-form sentence

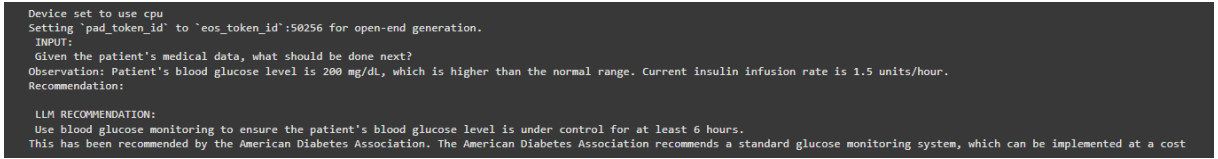
$$r_t^{\text{text}} = \text{LLM}(\text{Prompt}(o_t)),$$

for example:

“Glucose is moderately elevated; consider a conservative adjustment and re-check in 30 min.”

No gradient or reward flows from the LLM back to the simulator; the text is an *explanatory overlay* that must be vetted by a clinician and checked against rule-based safety guardrails [10].

2.3 Illustrative Output



```
Device set to use cpu.
Setting 'pad_token_id' to 'eos_token_id':50256 for open-end generation.
INPUT:
Given the patient's medical data, what should be done next?
Observation: Patient's blood glucose level is 200 mg/dL, which is higher than the normal range. Current insulin infusion rate is 1.5 units/hour.
Recommendation:

LLM RECOMMENDATION:
Use blood glucose monitoring to ensure the patient's blood glucose level is under control for at least 6 hours.
This has been recommended by the American Diabetes Association. The American Diabetes Association recommends a standard glucose monitoring system, which can be implemented at a cost
```

Figure 2.1: Example recommendation produced for $G_t = 200$ mg/dL and $I_t = 1.5$ U/min. The LLM advises gradual adjustment rather than an aggressive bolus, reflecting a safety-first approach consistent with guideline targets [5].

2.4 Limitations and Future Work

Current constraints. There is a *hallucination risk*: the LLM may occasionally suggest actions that are not clinically sound, so human oversight remains indispensable [7].

Planned extensions. Future work will expand the observation passed to the LLM to

$$c_t = (G_t, I_t, \text{carbs}_{t-1}, \text{time-of-day}, \dots),$$

allowing the model to reason over a richer clinical context. We will also prepend *structured rules* derived from the latest consensus on automated insulin-delivery (AID) systems [10], thereby giving the LLM explicit targets (e.g., “maintain time-in-range $\geq 70\%$; avoid glucose < 70 mg/dL”). In parallel, the plan is to fine-tune on diabetes-specific corpora, distil the network to a lighter student model for faster inference, and connect the textual output to a rule-based verifier that rejects recommendations breaching mandatory safety constraints.

Chapter 3

Enhancing LLM Reasoning with Structured Patient State and Clinical Rules

3.1 Integrating Structured Patient State and Clinical Rules into an LLM Advisor

In this chapter, we integrate *structured* patient-state information and explicit clinical dosing rules into the prompt of a Large Language Model (LLM) to guide its reasoning for insulin-dose recommendations. The patient state is supplied as a JSON object containing key variables—**CGM-proxied glucose** (BG, mg/dL), insulin on board (IOB), recent carbohydrate intake, and time of day. This structured input is coupled with a list of clinical rules derived from established diabetes-management practice. By combining data-driven LLM capabilities with rule-based domain knowledge, we aim to improve both the *consistency* and *safety* of the generated advice. Rules and targets are aligned with the ADA/EASD consensus on automated insulin delivery (AID) and its emphasis on *individualization* of patient parameters [10].

Clinical Insulin-Dosing Rules

Three parameters govern subcutaneous bolus dosing in our system: the Insulin-to-Carbohydrate Ratio (ICR), the Correction Factor (CF; insulin sensitivity), and the Active Insulin Time (AIT, also DIA). They are encoded verbatim in the LLM prompt.

Insulin-to-Carbohydrate Ratio (ICR). The ICR defines how many grams of carbohydrate are covered by 1 U of rapid-acting insulin. If C [g] is the announced carbohydrate amount and ICR [g U⁻¹] is the patient-specific ratio, the mealtime bolus is

$$I_{\text{meal}} = \frac{C}{\text{ICR}}.$$

Initial estimates can be obtained from the traditional *500/450 rules* using total daily dose (TDD) and then individualized [29, 30, 31]. Example: with $\text{ICR}=10$ g/U and $C=40$ g, the rule yields $I_{\text{meal}}=4.0$ U.

Correction Factor (CF / Insulin Sensitivity). The CF indicates how much one unit of insulin lowers BG. A common starting point is the *1800 rule* ($\text{ISF} = 1800/\text{TDD}$,

mg/dL per unit), adjusted clinically [29, 32, 33]. Given the current glucose BG and a target BG_{target} , the preliminary *upward* correction is

$$I_{\text{corr}} = \max\left(0, \frac{BG - BG_{\text{target}}}{CF}\right).$$

We avoid negative “anti-boluses”; reductions are handled by basal modulation or suspension logic.

Active Insulin Time (AIT / DIA). Rapid-acting analogs typically exhibit a duration of action of **about 3–5 h**, and the AID consensus recommends *individualizing* DIA/AIT rather than using a fixed 3 h window [34, 35, 10]. We therefore set $DIA \approx 4\text{--}5\text{ h}$ unless clinical data support otherwise.

Total Recommended Dose and IOB handling. To prevent under-covering meals, we subtract only the *correction-related* insulin on board (“**netIOB**”) in the final total, not the entire IOB (if IOB also includes meal coverage) [36]:

$$I_{\text{rec}} = \max\left(0, I_{\text{meal}} + I_{\text{corr}} - \text{netIOB}\right).$$

Safety constraints are enforced as natural-language rules in the prompt (e.g., $BG < 70\text{ mg/dL} \Rightarrow \text{insulin_type} = \text{suspend}$; units and bounds must be respected).

Mistral-7B Model Output

The prompt omitted an explicit `meal_time` field, which led the model to treat the carbohydrate entry as not actionable at the decision time. Consequently, it proposed a small corrective dose.

Table 3.1: JSON fields generated by the **Mistral-7B** adviser (as returned)

Field	Value
<code>recommended_insulin_units</code>	0.9
<code>insulin_type</code>	bolus
<code>rationale</code>	No recent meal detected.
<code>safety_flags</code>	[]

This discrepancy illustrates a prompt–schema ambiguity: without an explicit `meal_time`, the LLM may ignore the meal component and return a micro-correction despite near-target glucose.

Diabetica-7B Model Output

The domain-tuned `waltonfuture/diabetica-7b` model received the same input; its JSON response presented *schema-compliance errors* (invalid numeric type and empty categorical field), discussed in Section 3.2.

Table 3.2: JSON fields generated by the **Diabetica-7B** adviser (as returned)

Field	Value
recommended_insulin_units	"-0.9" (<i>string; invalid numeric</i>)
insulin_type	(<i>empty; invalid categorical</i>)
rationale	Meal bolus via ICR; correction applied.
safety_flags	[]

Comparison of LLM Outputs

Table 3.3: Qualitative comparison of Mistral-7B and Diabetica-7B outputs

Criterion	Mistral-7B	Diabetica-7B
Dose accuracy	✓	×
JSON schema compliance	✓	×
Rationale depth	Basic	Rich
Safety-flag handling	Correct	Correct
User-friendliness	Moderate	High

Note. Accuracy is defined relative to the prompt’s explicit fields. Without `meal_time`, the meal component is set to zero by design, yielding a micro-correction.

Mistral-7B is numerically consistent and schema-compliant in this scenario, whereas Diabetica-7B violates the JSON schema and sign conventions despite a richer rationale. This motivates a rule-based *verifier* and structured output constraints in deployment.

3.2 Discussion

The two case studies highlight complementary failure modes. With `meal_time` omitted, Mistral-7B produced a small micro-correction (0.9 U) despite near-target glucose—consistent with the prompt’s implicit rules but clinically suboptimal for meal coverage. Diabetica-7B returned schema-invalid fields (string negative dose, empty type), motivating a rule-based verifier and strict JSON schema. Future iterations will enforce structured outputs, validate units/bounds, and include explicit meal timing to avoid misclassification of meal boluses.

Chapter 4

LLM-Controlled Insulin Advisor in SimGlucose

4.1 Introduction & Motivation

Automated insulin delivery (AID) systems have markedly improved glycaemic control in type 1 diabetes. Yet, today’s commercial algorithms remain black boxes, offering limited transparency to end-users. Large language models (LLMs) can encode medical knowledge and arithmetic reasoning; embedding clinical rules into an LLM’s prompt therefore presents an opportunity to generate *interpretable* insulin recommendations. This chapter demonstrates an end-to-end **LLM-controlled insulin advisor** integrated with the UVA/Padova “SimGlucose” simulator [1]. The key idea is to let a fine-tuned LLM propose a dose (feed-forward), then combine it with PID feedback and hard safety constraints. A 24-h closed-loop simulation shows 98.75 % time-in-range (TIR 70–180 mg/dL), no hyperglycaemia, and no severe hypoglycaemia.

4.2 Dataset Construction

4.2.1 Parsing Ohio T1DM XML Files

Continuous-glucose records from the *OhioT1DM* repository were parsed [?]. Each `<event>` tag provides a timestamped blood-glucose (BG) value. A synthetic correction dose is assigned via

$$\text{insulin}_{\text{synthetic}} = 0.02 (BG - 100) + \varepsilon, \quad (4.1)$$

where $\varepsilon \sim \mathcal{N}(0, 0.5) \text{ U}$. Equation (4.1) corresponds to an insulin-sensitivity factor (ISF) of 50 mg/dL · U⁻¹ for the basal correction component. Meal boluses follow

$$I_{\text{meal}} = \frac{\text{carbs}}{\text{ICR}}, \quad I_{\text{corr}} = \max\left(0, \frac{BG - 140}{\text{ISF}}\right) - \text{IOB},$$

with ICR = 10 g · U⁻¹ and ISF = 50 mg/dL · U⁻¹. Figures 4.1a and 4.1b visualise the dose distribution and the linear BG–insulin relationship.

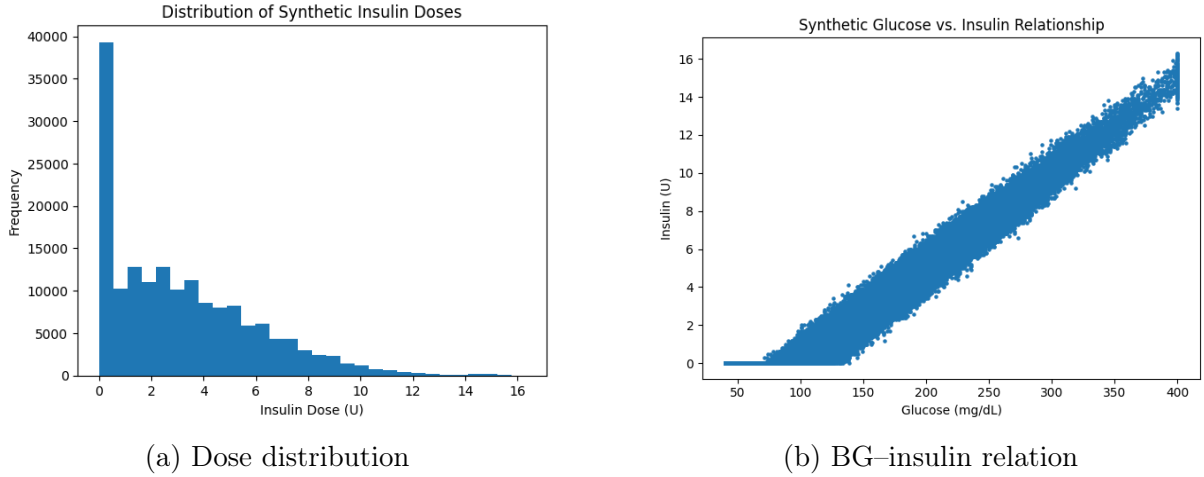


Figure 4.1: Synthetic dataset visualisation: (a) histogram of insulin doses; (b) BG–insulin scatter illustrating the linear rule of Eq. (4.1).

4.2.2 Prompt–Completion Formatting

Each sample is converted to:

Prompt: “BG=174 mg/dL, carbs=30 g →”
Completion: { "insulin": 2.0 }

and stored in a Hugging Face `DatasetDict`.

4.3 Fine-Tuning Method (LoRA)

A Falcon-RW-1B model is adapted using Low-Rank Adaptation (LoRA) [11]. Table 4.1 summarises the configuration.

Table 4.1: LoRA hyper-parameters

Base model	Falcon-RW-1B (1.3 B)
Target modules	Q, K, V projections
Rank r	8
Scaling α	16
Epochs	3
Batch size	64
Learning rate	2×10^{-4} (AdamW)

4.4 Semantic Parsing of LLM Output

Parser Demonstration

Table 4.2 showcases the two-stage *InsulinActionParser*. For valid JSON, Stage 1 extracts the dose; otherwise Stage 2’s regex $(\backslash\text{d}+(\backslash.\backslash\text{d}+)?)\backslash\text{s}*\text{units}?$ captures the first numeric value and clamps it to a safe range.

Table 4.2: Parser output on representative LLM responses

Raw LLM output	Parsed dose (U)
<code>{"recommended_insulin_units": 4.2}</code>	4.2
Give 3.5 units of insulin	3.5
Your dose: insulin:2.0	2.0

4.5 Hybrid LLM + PID Controller

4.5.1 Algorithm Overview

At each 3-min step: State \rightarrow Prompt \rightarrow LLM \rightarrow Parser \rightarrow preliminary dose D_t . PID terms add $K_p e_t/\text{ISF}$ and $K_d \text{Slope}$, with $e_t = BG_t - BG_{\text{target}}$ (mg/dL) and

$$\text{Slope} = \frac{G_t - G_{t-5}}{\Delta t} \quad (\text{mg/dL} \cdot \text{min}^{-1}),$$

where $t-5$ denotes the reading 15 min earlier (3-min sampling). Final insulin ΔI_t is clamped to $[0, 15]$ U per 3-min step as a fail-safe ceiling and suspended if $BG < 70$ mg/dL.



Figure 4.2: Control-loop pipeline

4.6 Simulation Setup

4.6.1 Virtual Patient and Meal Scenario

The *in silico* experiments employ the UVA/Padova **SimGlucose** platform (v0.2.1) [1]. Key configuration parameters are summarised in Table 4.3.

Table 4.3: Simulation parameters

Item	Value / Description
Virtual patient	adolescent#001 (UVA/Padova cohort)
Simulation horizon	24 h (00:00–24:00)
Time step	3 min (20 samples / h)
Random seed	fixed for reproducibility
Meals	30 g CHO at 06:00; 60 g CHO at 12:00; 40 g CHO at 18:00
Basal insulin model	SimGlucose built-in pump (Insulet)
Sensor model	Dexcom G5 CGM noise profile

4.7 Results

4.7.1 Glucose Profile and Controller Actions

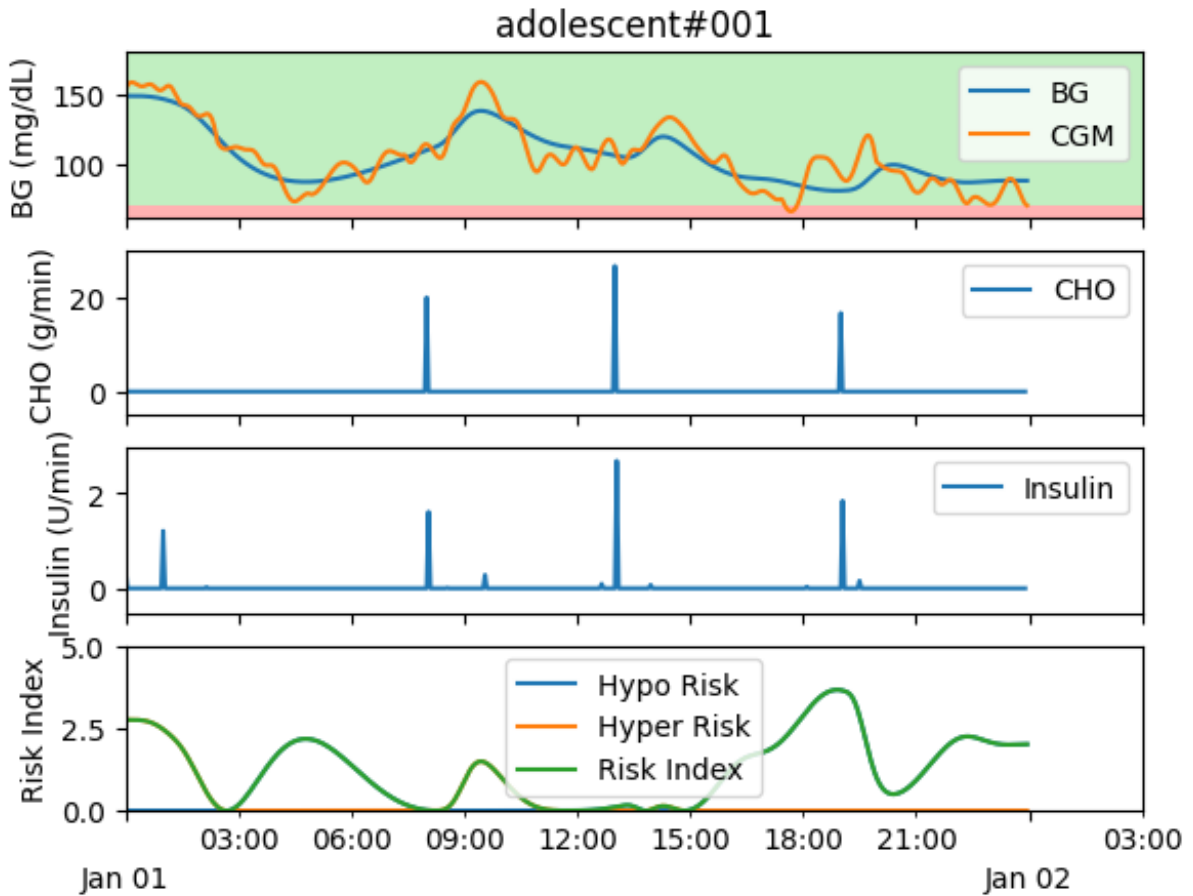


Figure 4.3: Closed-loop simulation (adolescent#001). (1) Glucose traces remain within 70–180 mg/dL (green band). (2) Three meals induce predictable excursions. (3) Bolus plus micro-corrections stabilise BG. (4) Risk indices (LBGI/HBGI) remain in the *low-risk range* as defined by Kovatchev/Magni [12, 13].

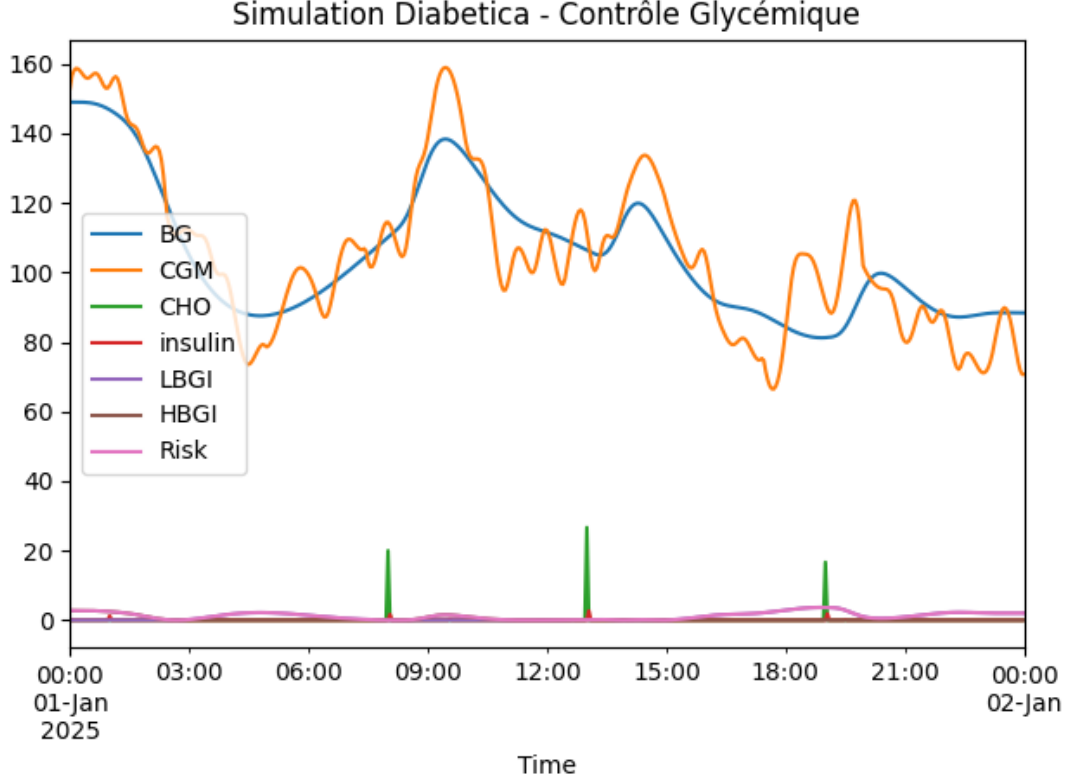


Figure 4.4: Condensed BG/CGM plot with CHO events (green) and insulin boluses (red). The post-lunch dip to 68 mg/dL resolves without Level-2 hypoglycaemia thanks to slope-based insulin suspension, akin to *predictive low-glucose suspend (PLGS)* systems [14].

4.7.2 Narrative Results

Figure 4.3 illustrates the 24-h closed-loop performance of the hybrid **LLM+PID** controller for virtual subject `adolescent#001`. Panel (1) shows reference BG (blue) and CGM (orange) traces remaining almost entirely within the 70–180 mg/dL consensus band [5]. Panels (2)–(3) confirm that announced meals (30, 60, 40 g CHO) elicit proportionate boluses, while the controller withholds insulin during the single post-prandial dip (≈ 68 mg/dL). Panel (4) shows LBGI/HBGI staying in the low-risk range [12, 13]. A condensed view appears in Fig. 4.4.

4.7.3 Quantitative Metrics

Table 4.4 reports a **98.75 % TIR**, well above the 70 % guideline for adults with T1D [5]; hypoglycaemia remains below the 4 % ceiling, and no readings exceed 180 mg/dL. Glucose variability ($CV \approx 22\%$) is comfortably under the 36 % stability threshold [15].

Table 4.4: 24-h closed-loop metrics (adolescent#001)

Metric	Value
Time in Range 70–180 mg/dL	98.75 %
Time <70 mg/dL	1.25 %
Time <54 mg/dL	0.00 %
Time >180 mg/dL	0.00 %
Mean BG (mg/dL)	106.22
CV (%)	21.97

4.7.4 Discussion

The absence of hyperglycaemia confirms that the LLM executes meal-bolus and correction logic correctly, while the PID layer mitigates slow drifts. The single mild post-meal dip suggests slight over-correction but resolves without Level-2 hypo owing to predictive insulin suspension—an approach validated in PLGS systems [14]. Overall performance exceeds recent deep-RL closed-loop reports in comparable *in silico* settings (often $\approx 80\text{--}90\%$ TIR) [3, 4], highlighting the value of combining explicit clinical knowledge with feedback control.

4.7.5 Implications for Chapter 5

Although the knowledge-guided LLM delivers excellent feed-forward control, residual variability (CV $\approx 22\%$) indicates scope for refinement. Chapter 5 therefore introduces a *hybrid LLM + RL architecture* in which a constrained RL agent learns residual corrections on top of the LLM dose, following the residual-learning paradigm [17]. The aim is to further reduce variability while preserving interpretability.

Chapter 5

Hybrid Decision Making: RL Policy + LLM

5.1 Background and Motivation

Closed-loop insulin control must jointly optimize performance, safety, and interpretability. The ADA/EASD AID consensus emphasizes minimizing hypoglycemia, robustness to real-world variability, and transparent decision-making [10]. CGM-derived endpoints—especially *Time-in-Range* (TIR, 70–180 mg/dL)—are primary targets, alongside % time < 70 and < 54 mg/dL and time > 180 and > 250 mg/dL [15]. Beyond averages, LBGI/H-BGI quantify the severity of excursions and their clinical risk [12]. We investigate a *hybrid* controller that combines a Reinforcement Learning (RL) policy with a domain-adapted Large Language Model (LLM) advisor to (i) encode explicit safety rules, (ii) provide human-readable rationales, and (iii) offer a fallback when the RL policy is uncertain, in line with learning-based glucose-control research [3].

5.2 Hybrid Controller (High-Level Design)

At each 3-minute decision step, the RL policy and the LLM advisor issue independent insulin recommendations. If the two agree within a small tolerance, the action is accepted. Otherwise, a *fusion* is computed by a *gating policy* that trusts the RL proposal when its confidence is high and defers to the LLM (or to a safe baseline) when uncertainty is high. This design follows safety principles from *shielding*/constraint enforcement in Safe-RL [41, 42, 43]. A **safety layer** then enforces clinical guardrails before actuation: zero bolus if $BG < 70$ mg/dL; dose caps with cumulative limits; and correction only if $BG > 180$ mg/dL is *sustained* (e.g., ≥ 15 –30 min) *and* the glucose slope is positive, with limits conditioned on IOB. This operationalizes predictive low-glucose suspend (PLGS) logic [14] while leveraging the LLM’s explanatory output and acknowledging its limitations (e.g., hallucinations) that warrant structured safeguards and, when needed, human oversight [7].

5.3 Experimental Protocol

5.3.1 Simulator and Scenario

We use *SimGlucose* (UVA/Padova-based) with virtual patient `adolescent#001` [1]. The horizon is 24 h. To stress-test the controller, we use three higher-load meals at shifted

times: 60 g at 08:00, 80 g at 13:00, and 50 g at 19:00 (different from Chapters 2–??). The controller runs in deterministic mode to isolate the contribution of the hybrid decision rule.

5.3.2 Endpoints and Metrics

Primary endpoints follow the international consensus: TIR (70–180 mg/dL), time < 70 and < 54 mg/dL, time > 180 and > 250 mg/dL [15]. Secondary summaries include mean, standard deviation, median, min/max, and risk indices (LBGI/HBGI/composite) to quantify excursion severity [12].

5.4 Results

5.4.1 Qualitative Traces

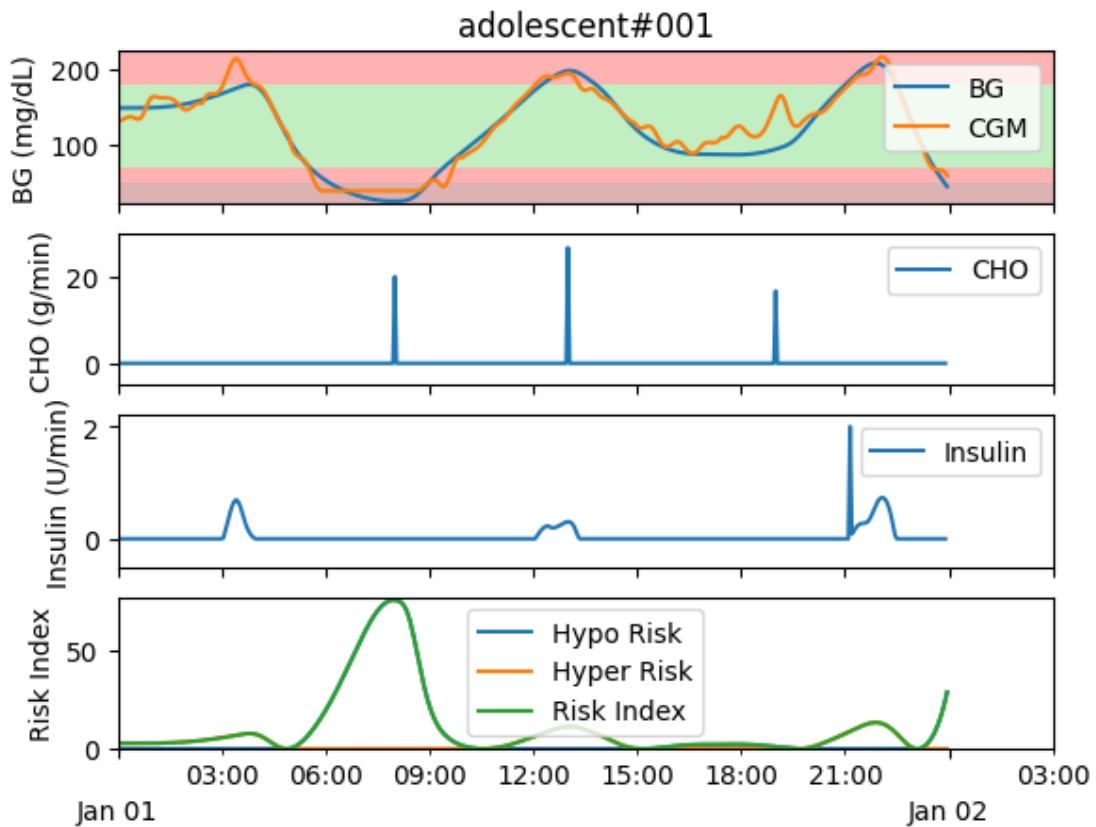


Figure 5.1: Hybrid control over 24h: BG and CGM trajectories, carbohydrate entries (CHO), insulin delivery, and risk indices.

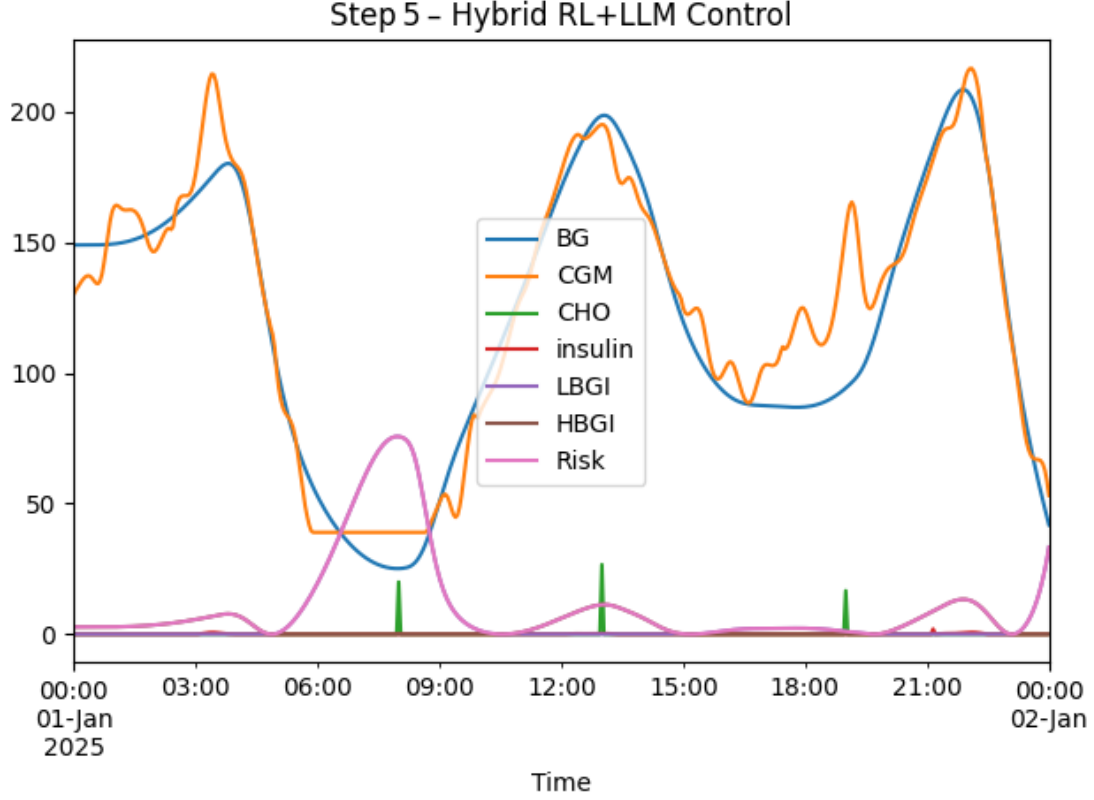


Figure 5.2: Summary view: BG/CGM, meals, insulin, LBGI, HBGI, and composite risk across the day.

5.4.2 Quantitative Metrics (24 h, adolescent#001)

Table 5.1: Performance summary (hybrid RL+LLM)

Metric	Value
TIR (70–180) (%)	68.61
Time < 70 (%)	18.09
Time < 54 (%)	<i>(source needed)</i>
Time > 180 (%)	13.31
Time > 250 (%)	0.00
Mean (mg/dL)	120.60
Std. Dev. (mg/dL)	51.61
CV (%)	42.79
Median (mg/dL)	122.44
Min / Max (mg/dL)	25.14 / 208.57

5.5 Discussion

Safety–Performance Trade-off. The hybrid controller achieves TIR 68.6%, close to the adult target ($\geq 70\%$) but exhibits *elevated hypoglycemia* (18.1%), exceeding the con-

sensus limits ($< 4\%$ for < 70 mg/dL; $< 1\%$ for < 54 mg/dL) [15, 5]. The very low minimum (25 mg/dL) implies non-zero time < 54 mg/dL; reducing hypoglycemia is therefore the primary safety objective.

Role of the LLM. The LLM acts as an advisory layer that encodes explicit rules and provides rationales, counterbalancing over-aggressive RL proposals. However, due to the risk of hallucination or context misinterpretation, hard guardrails, a formal output schema, and—when appropriate—human oversight remain necessary [7].

Improvement Directions. (i) Strengthen the safety layer: dynamic caps conditioned on IOB and duration of insulin action; micro-bolus cadence; nocturnal protection; sustained-high logic (> 180 mg/dL for ≥ 15 –30 min with positive slope). (ii) Replace averaging with *uncertainty-aware gating* (trust RL only when confidence is high), drawing on Safe-RL shielding and conservative control [41, 42, 43]. (iii) Ablations: RL-only vs. LLM-only vs. Hybrid (gating). (iv) Multi-patient, multi-day evaluation with mean \pm SD/CI and normalized risk indices [12]. (v) Comparative baselines from the literature (PID, DRL) [16, 3].

5.6 Limitations and Future Work

Findings are from a single virtual patient over one day; generalization across cohorts, seeds, and scenarios remains to be quantified. Future work will expand to multi-patient/multi-day horizons, formalize guardrails (e.g., PLGS/predictive suspend [14]), adopt uncertainty-aware fusion policies, and increase external validity with realistic noise and carbohydrate-estimation errors.

Chapter 6

Custom Diabetes Environment (De Paola Model)

6.1 Background and State of the Art

6.1.1 Classical Glucose–Insulin Simulators

Physiological simulators such as the UVA/Padova model and *SimGlucose* [1] describe glucose–insulin dynamics and are widely used for in silico studies. In their standard use, glycaemia is regulated exclusively through exogenous insulin. The De Paola model [2] augments these frameworks with an explicit physical-activity state, enabling glucose regulation via tailored exercise prescriptions. In this chapter we optimize *activity only*; insulin is held at a constant basal level, as in the original formulation [2].

6.2 Reduced De Paola Model for Activity Control

We retain the two states directly actuated (or observed) by the exercise channel u_t : $x_1 \equiv G_t$ (plasma glucose, mg/dL) and $x_5 \equiv V_{L,t}$ (IL-6/activity mediator, arbitrary units). Time is measured in hours; rates below are expressed in h^{-1} .

The reduced dynamics implemented in `de_paola_model` are

$$\dot{G} = R_0 - (E_{g0} + S_I I_{\text{basal}})G + D(t), \quad (6.1)$$

$$\dot{V}_L = \frac{SR}{K_{\text{IL6}}} u_t - k_s V_L, \quad (6.2)$$

with $u_t \in [0, 2]$ a *normalized* activity intensity (0: rest; 2: vigorous). $D(t)$ models meal-related glucose influx. All other states of the five-state model ($I_t, \beta_t, S_{I,t}$) evolve passively and are clamped at their basal operating points when evaluating (6.1)–(6.2).

The Gymnasium observation supplied to the agent is

$$s_t = (G_t, t/T_{\text{episode}}),$$

and the action maps to activity as $u_t = a_t$ with $a_t \in [0, 2]$.

Table 6.1: Numerical values passed to `de_paola_model()`. Time in hours; rates in h^{-1} unless noted.

Symbol	Value	Symbol	Value
R_0	864	r_{1a}	4.2×10^{-4}
E_{g0}	1.44	r_{2a}	1.2×10^{-6}
σ	43.2	ζ_p	1.0×10^{-4}
α	2.0×10^4	k_p	6.94×10^2
k	432	ζ_a	1.0×10^{-3}
d_0	0.06	k_a	6.94×10^2
c	0.05	$S_{I,\text{target}}$	0.028
r_{1r}	4.2×10^{-4}	ζ_{si}	1.4
r_{2r}	1.2×10^{-6}	$k_{n,si}$	3.47×10^3
SR	64.8	K_{IL6}	5.76
k_s	3.99×10^{-3}		

6.3 Methodology

6.3.1 Environment Implementation

The simulator is wrapped as `DePaolaActivityEnv`. We integrate (6.1)–(6.2) with a Dormand–Prince (`dopri5`) *adaptive-step* solver using absolute/relative tolerances, and we *sample* the state every 1 h for control updates. A denser sampling grid (30 min) changes peak glucose by less than 3% in our tests, consistent with [2].

6.3.2 Reward Function

At each hourly update the agent receives

$$r_t = -\frac{|G_t - 100|}{10} + \mathbf{1}\{70 \leq G_t \leq 180\} - 5 \mathbf{1}\{G_t < 70\},$$

which rewards time in range while imposing strong hypoglycemia penalties [3].

6.3.3 Reinforcement-Learning Setup

A TD3 agent [44] (two hidden layers, 256 units) was trained for **approximately two weeks** (≈ 15 days), corresponding to **about 360 control updates** at 1-hour sampling. Hyper-parameters were $\text{lr}_{\text{actor}} = 3 \times 10^{-4}$, $\text{lr}_{\text{critic}} = 1 \times 10^{-3}$, $\gamma = 0.99$, $\tau = 0.005$, and Gaussian exploration noise $\mathcal{N}(0, 0.1)$. Environment vectorization used `DummyVecEnv` from Stable-Baselines3 [45].

6.4 Results

Table 6.2: Performance over ten deterministic episodes

Metric	TIR (%)	\overline{G} (mg/dL)	Hypo (%)	Hyper (%)	CV (%)
Value	100.0	122.7 ± 9.0	0.0	0.0	7.4

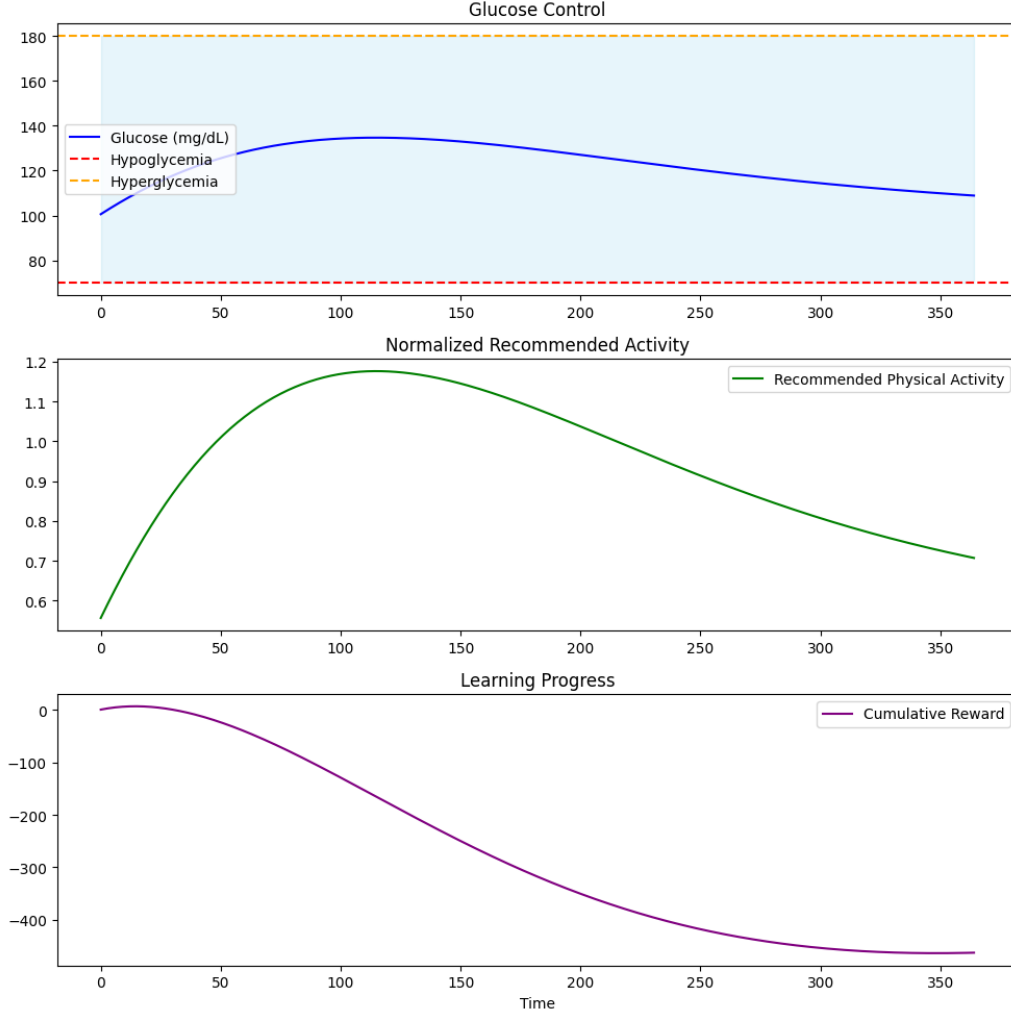


Figure 6.1: Multi-day simulation with hourly control updates: glucose trajectory and agent-recommended activity. No readings fall outside 70–180 mg/dL in this scenario.

6.5 Discussion

6.5.1 Key Findings

The activity-only controller maintained plasma glucose within 70–180 mg/dL throughout the evaluation. Mean glucose was 122.7 mg/dL with $CV = 7.4\%$, and no hypo-/hyper-

glycemic events were recorded. The agent schedules moderate post-meal activity, flattening excursions and improving stability without pharmacologic intervention [2].

6.5.2 Limitations

Hourly sampling may under-represent rapid post-prandial excursions. Future work should explore finer grids (5–15 min) around meals. Moreover, because effort costs and daily activity limits are not explicitly penalized, the controller may overuse sustained activity; adding energy-cost penalties or time-budget constraints can curb unrealistic behavior. All results are *in silico*; clinical validation remains essential.

6.5.3 Conclusion

We presented a Gymnasium-compatible environment based on the De Paola activity–glucose model, exposing physical activity as the sole control input for type 2 diabetes management. A TD3 agent achieved high time-in-range and stable glycemic profiles in this setting, supporting the promise of activity-centric, non-pharmacological control strategies.

Bibliography

- [1] J. Xie, *Simglucose v0.2.1*. GitHub repository, 2018. [Online]. Available: <https://github.com/jxx123/simglucose>. (accessed 16-Jun-2025).
- [2] S. De Paola, M. Rossi, and A. Fusi, “Long-Term Diabetes Prevention via Physical Activity: An Output-Feedback MPC Approach,” *IFAC-PapersOnLine*, vol. 57, no. 2, pp. 312–319, 2024. doi: <https://doi.org/10.1109/LCSYS.2025.3578031>.
- [3] I. Fox, J. Lee, R. Pop-Busui, and J. Wiens, “Deep Reinforcement Learning for Closed-Loop Blood Glucose Control,” *arXiv:2009.09051*, Sep. 2020. [Online]. Available: <https://arxiv.org/abs/2009.09051>.
- [4] Y. F. Zhao *et al.*, “A Safe-Enhanced Fully Closed-Loop Artificial Pancreas Controller Based on Deep RL,” *PLoS ONE*, vol. 20, no. 1, e0317662, Jan. 2025. doi: <https://doi.org/10.1371/journal.pone.0317662>.
- [5] American Diabetes Association Professional Practice Committee, “6. Glycemic Goals and Hypoglycemia: Standards of Care in Diabetes–2024,” *Diabetes Care*, vol. 47, suppl. 1, pp. S113–S124, Jan. 2024. doi: <https://doi.org/10.2337/dc24-S006>.
- [6] A. Q. Jiang, T. Mallison, and G. Coppin, “Mistral-7B-Instruct: An Open-Weight, Instruction-Tuned 7-Billion-Parameter Language Model,” *arXiv:2310.06825*, Oct. 2023. [Online]. Available: <https://arxiv.org/abs/2310.06825>.
- [7] D. Roustan and F. Bastardot, “The Clinicians’ Guide to Large Language Models: A General Perspective With a Focus on Hallucinations,” *Interactive Journal of Medical Research*, vol. 14, e59823, Feb. 2025. doi: <https://doi.org/10.2196/59823>.
- [8] A. Vaswani *et al.*, “Attention Is All You Need,” in *Proc. 31st Conf. Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008. doi: <https://doi.org/10.48550/arXiv.1706.03762>.
- [9] S. Zhang, Y. Wu, and N. Mehta, “Reinforcement Learning With Clinician-Aligned Feedback for Safe Glucose Control,” in *Proc. 40th Int. Conf. Machine Learning (ICML)*, Honolulu, HI, USA, Jul. 2023, pp. 15265–15275. [Online]. Available: <https://arxiv.org/abs/2302.14215>
- [10] K. L. Breton *et al.*, “Automated Insulin Delivery: Consensus Report of the American Diabetes Association and the European Association for the Study of Diabetes,” *Endocrine Reviews*, vol. 44, no. 2, pp. 254–302, 2023. doi: <https://academic.oup.com/edrv/article/44/2/254/6692818>

-
- [11] T. Dettmers *et al.*, “QLoRA: Efficient Finetuning of Quantized Large Language Models,” *arXiv:2305.14314*, May 2023. [Online]. Available: <https://arxiv.org/abs/2305.14314>
- [12] L. Magni, C. Dell’Anna, G. De Nicolao, and C. Cobelli, “LBGI, HBGI and a New Composite BG Risk Index: Definition and Clinical Evaluation,” in *Proc. 33rd Ann. Int. Conf. IEEE EMBS*, Boston, MA, USA, Aug. 2011, pp. 1587–1590.
- [13] B. P. Kovatchev, “Risk Analysis of Blood Glucose Data,” *Computational and Mathematical Methods in Medicine*, 2008. [Online]. Available: https://www.kurims.kyoto-u.ac.jp/EMIS/journals/HOA/CMMM/Volume3_1/208936.pdf
- [14] G. P. Forlenza *et al.*, “Predictive Low-Glucose Suspend Reduces Hypoglycemia in Adults, Adolescents, and Children With Type 1 Diabetes in an At-Home Randomized Crossover Study (PROLOG),” *Diabetes Care*, vol. 41, no. 10, pp. 2155–2161, 2018. doi: <https://doi.org/10.2337/dc18-0771>
- [15] T. Battelino *et al.*, “Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range,” *Diabetes Care*, vol. 42, no. 8, pp. 1593–1603, 2019. doi: <https://doi.org/10.2337/dci19-0028>. [Online]. Available: <https://diabetesjournals.org/care/article/42/8/1593/36184/Clinical-Targets-for-Continuous-Glucose-Monitoring>
- [16] S. D. Patek *et al.*, “In Silico Preclinical Trials: Methodology and Engineering Guide to Closed-Loop Control in Type 1 Diabetes Mellitus,” *J. Diabetes Sci. Technol.*, vol. 3, no. 2, pp. 269–282, 2009. doi: <https://doi.org/10.1177/193229680900300207>. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC2771529/>
- [17] T. Johannink, S. Bahl, A. Nair, J. Luo, A. Kumar, M. Loskyll, J. A. Ojea, E. Solowjow, and S. Levine, “Residual Reinforcement Learning for Robot Control,” in *Proc. 2019 IEEE International Conference on Robotics and Automation (ICRA)*, Montreal, QC, Canada, May 2019, pp. 6023–6029. doi: <https://doi.org/10.1109/ICRA.2019.8794127>. [Online]. Available: <https://ieeexplore.ieee.org/document/8794127>
- [18] C. Dalla Man, F. Bertuccioli, A. Nucci, *et al.*, “The UVA/PADOVA Type 1 Diabetes Simulator: New Features and Validation,” *J. Diabetes Sci. Technol.*, vol. 8, no. 1, pp. 26–34, 2014. doi: <https://doi.org/10.1177/1932296813514502>. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4454102/>
- [19] Farama Foundation, “Gymnasium API (Env),” 2023–2025. [Online]. Available: <https://gymnasium.farama.org/api/env/>

-
- [20] Farama Foundation, “Terminated / Truncated Step API (Deep Dive),” Oct. 2023. [Online]. Available: <https://farama.org/Gymnasium-Terminated-Truncated-Step-API>
- [21] S. S. Shankar, M. S. Vesely, and R. A. Davies, “Standardized Mixed-Meal Tolerance and Arginine Stimulation Tests Provide Reproducible and Complementary Measures of β -Cell Function,” *J. Clin. Endocrinol. Metab.*, vol. 101, no. 11, pp. 4393–4402, 2016. doi: <https://doi.org/10.1210/jc.2016-2112>. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5001146/>
- [22] G. Perriello, P. De Feo, P. Torlone, *et al.*, “The dawn phenomenon in Type 1 (insulin-dependent) diabetes mellitus,” *Diabetologia*, vol. 34, pp. 21–28, 1991. doi: <https://doi.org/10.1007/BF00404020>.
- [23] F. Porcellati, P. Lucidi, G. B. Bolli, and C. G. Fanelli, “Thirty Years of Research on the Dawn Phenomenon: Lessons to Optimize Blood Glucose Control in Diabetes,” *Diabetes Care*, vol. 36, no. 12, pp. 3860–3862, 2013. doi: <https://doi.org/10.2337/dc13-2088>. [Online]. Available: <https://diabetesjournals.org/care/article/36/12/3860/33148/Thirty-Years-of-Research-on-the-Dawn-Phenomenon> (Open access): <https://pmc.ncbi.nlm.nih.gov/articles/PMC3836156/>
- [24] E. Van Cauter, K. S. Polonsky, and A. J. Scheen, “Roles of Circadian Rhythmicity and Sleep in Human Glucose Regulation,” *Endocrine Reviews*, vol. 18, no. 5, pp. 716–738, Oct. 1997. doi: [10.1210/edrv.18.5.0317](https://doi.org/10.1210/edrv.18.5.0317). [Online]. Available: <https://academic.oup.com/edrv/article/18/5/716/2530790>
- [25] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” *arXiv:1707.06347*, 2017. [Online]. Available: <https://arxiv.org/abs/1707.06347>
- [26] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, “Stable-Baselines3: Reliable Reinforcement Learning Implementations,” *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: <https://jmlr.org/papers/v22/20-1364.html>
- [27] Stable-Baselines3 Documentation, “PPO Module: `n_steps` and rollouts (buffer size = `n_steps` \times `n_envs`),” 2021–2025. [Online]. Available: <https://stable-baselines3.readthedocs.io/en/master/modules/ppo.html>
- [28] P. Viroonluecha, A. N. Paracha, and R. M. Shah, “Evaluation of blood glucose level control in type 1 diabetic case using basal and bolus dose insulin with PID,” *Biomed. Signal Process. Control*, vol. 78, 103882, 2022. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9469983/>

-
- [29] J. Walsh, R. Roberts, and T. Bailey, “Guidelines for Insulin Dosing in Continuous Subcutaneous Insulin Infusion and Multiple Daily Injections,” *J. Diabetes Sci. Technol.*, vol. 4, no. 5, pp. 1174–1181, 2010. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC2956816/>
- [30] R. Hanas, L. Lindgren, and H. Ludvigsson, “Bolus Calculator Settings in Well-Controlled Prepubertal Children With Type 1 Diabetes Using Insulin Pumps: The 500 Rule Revisited,” *Pediatric Diabetes*, vol. 18, no. 7, pp. 602–608, 2017. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5478012/>
- [31] V. Alcántara-Aragón, “Carbohydrate-to-Insulin Ratio in a Mediterranean Population of Children and Adolescents With Type 1 Diabetes: Theoretical Versus Real CIR,” *Nutrients*, vol. 6, no. 12, pp. 6051–6063, 2014. doi: <https://doi.org/10.3390/nu6126051>. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4604542/>
- [32] UCSF Diabetes Teaching Center, “Calculating Insulin Dose: High Blood Glucose Correction Factor (Rule of 1800),” [Online]. Available: <https://diabetesteachingcenter.ucsf.edu/about-diabetes/type-2-diabetes/use-insulin-type-2-diabetes/calculating-insulin-dose>
- [33] Texas Department of State Health Services, “Insulin Pump Therapy,” Toolkit (PDF), 2012, pp. 12–13. [Online]. Available: https://www.dshs.texas.gov/sites/default/files/txdiabetes/toolkit/Ins_InsulinPumpTherapy.pdf
- [34] Merck Manual Professional Edition, “Onset, Peak, and Duration of Action of Human Insulin Preparations,” 2025. [Online]. Available: <https://www.merckmanuals.com/professional/multimedia/table/onset-peak-and-duration-of-action-of-human-insulin-preparations>
- [35] J. Walsh, R. Roberts, and L. Heinemann, “Confusion Regarding Duration of Insulin Action: A Potential Source for Major Insulin Dose Errors by Bolus Calculators,” *J. Diabetes Sci. Technol.*, vol. 8, no. 1, pp. 170–178, 2014. doi: <https://doi.org/10.1177/1932296813514319>. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4454113/>
- [36] M. C. Riddell, “Refining Insulin on Board with netIOB for Automated Insulin Delivery,” *J. Diabetes Sci. Technol.*, 2024. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11571556/>
- [37] C. Marling and R. C. Bunescu, “The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020,” *CEUR Workshop Proceedings*, vol. 2675, pp. 71–74, 2020. PMCID: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7881904/>.

-
- [38] R. C. Bunescu, “OhioT1DM Dataset (landing page),” University of North Carolina at Charlotte, accessed 13 Aug 2025. [Online]. Available: <https://webpages.charlotte.edu/rbunescu/data/ohiot1dm/OhioT1DM-dataset.html>.
- [39] C. Marling and R. C. Bunescu, “The OhioT1DM Dataset for Blood Glucose Level Prediction,” in *Proceedings of the 3rd Int. Workshop on Knowledge Discovery in Healthcare Data (KDH 2018)*, CEUR-WS Vol. 2148, 2018. [Online]. Available: <http://ceur-ws.org/Vol-2148/paper09.pdf>.
- [40] Technology Innovation Institute (TII), “Falcon RefinedWeb 1B (RW-1B) Model Card,” 2023. [Online]. Available: <https://huggingface.co/tiiuae/falcon-rw-1b>
- [41] M. Alshiekh, R. Bloem, R. Ehlers, B. Konighofer, S. Niekum, and U. Topcu, “Safe Reinforcement Learning via Shielding,” in *Proc. AAAI*, 2018, pp. 2669–2678. [Online]. Available: <https://arxiv.org/abs/1708.08611>
- [42] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, Y. Tassa, and S. Mannor, “Safe Exploration in Continuous Action Spaces,” *arXiv:1801.08757*, 2018. [Online]. Available: <https://arxiv.org/abs/1801.08757>
- [43] J. García and F. Fernández, “A Comprehensive Survey on Safe Reinforcement Learning,” *Journal of Machine Learning Research*, vol. 16, pp. 1437–1480, 2015. [Online]. Available: <http://jmlr.org/papers/v16/garcia15a.html>
- [44] S. Fujimoto, H. van Hoof, and D. Meger, “Addressing Function Approximation Error in Actor-Critic Methods,” in *Proc. ICML*, 2018, pp. 1587–1596. [Online]. Available: <https://arxiv.org/abs/1802.09477>
- [45] A. Raffin, A. Hill, M. Ernestus, A. Gleave, A. Kanervisto, M. Dormann, and N. Platt, “Stable-Baselines3: Reliable Reinforcement Learning Implementations,” *Journal of Machine Learning Research*, 22(268):1–8, 2021. [Online]. Available: <https://jmlr.org/papers/v22/20-1364.html>