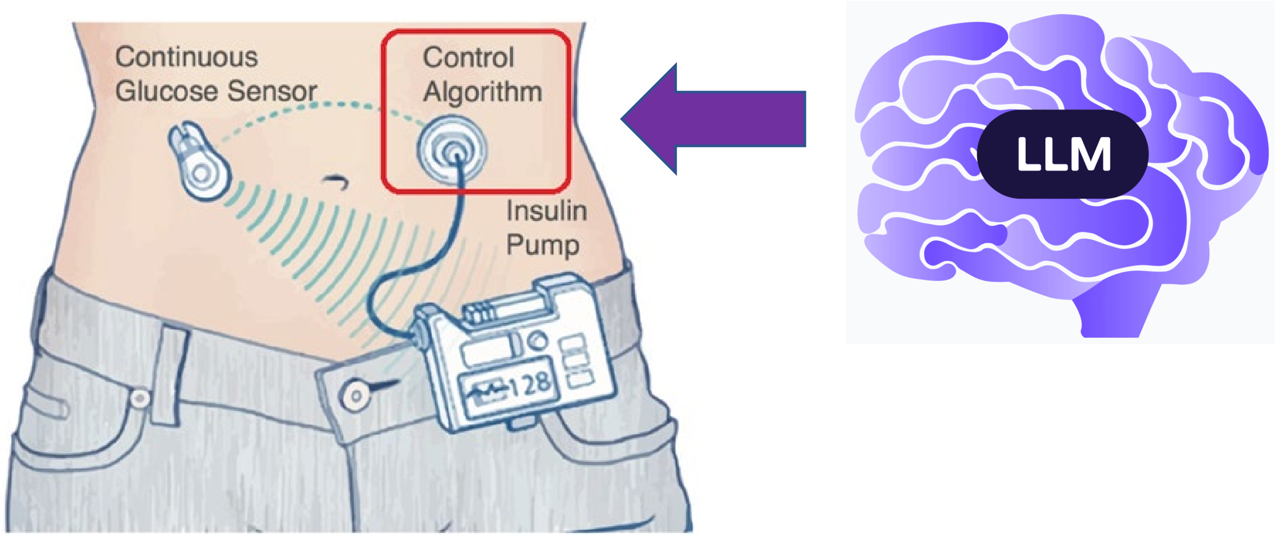
NeurIPS 2025 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences

Explainable Insulin Pump Control with LLMs for Type 1 Diabetes

NEURAL INFORMATION PROCESSING SYSTEMS 2025

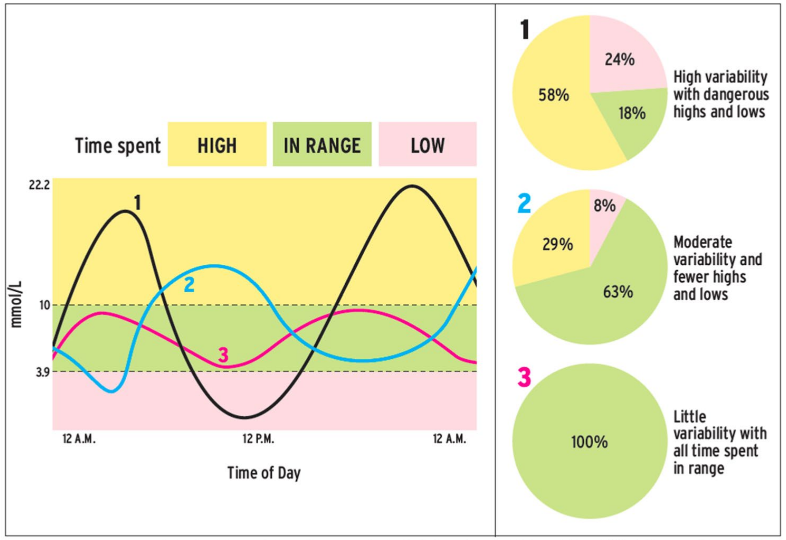
Maya Sarkar



Introduction: Trust Deficit in AI for Type 1 Diabetes Insulin Dosing &

The main goal of Type 1 diabetes management is to **maintain glucose levels within a healthy range**. (Time In Range (TIR)

- Patients use insulin management devices for continuous glucose monitoring (CGM) systems, to measure glucose levels every few minutes.
- Insulin dose from insulin pumps can be controlled by a patients' glucose level monitored by (CGM) device.



Adapted from Adam Brown, et al. diaTribe, August 2016. https://diatribe.org/BeyondA1c

The Problem: The Trust Deficit in AI Healthcare

- **Reinforcement Learning (RL)** can create powerful, automated insulin pumps that learn from data.
- **But they are "black boxes."** They can't explain *why* they make a life-critical decision.
- This creates a **trust deficit** for patients and clinicians, blocking the adoption of safer, more effective technology.

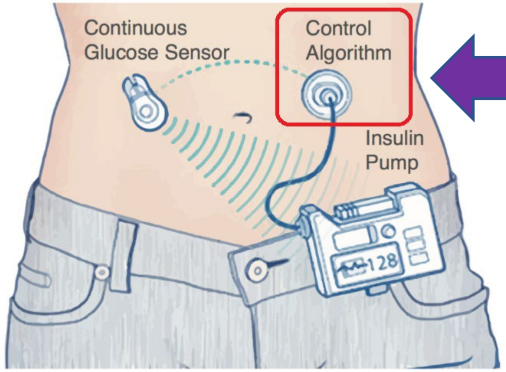
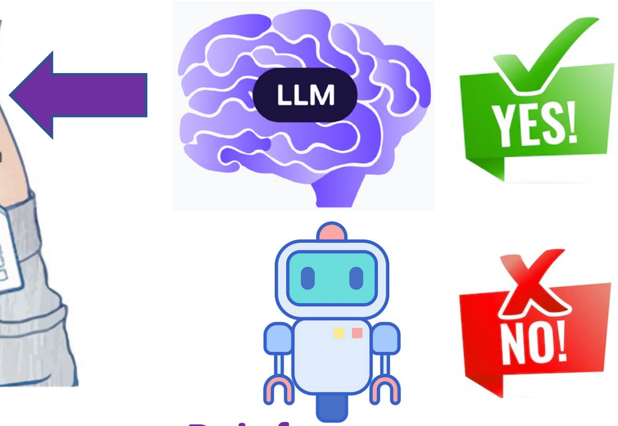
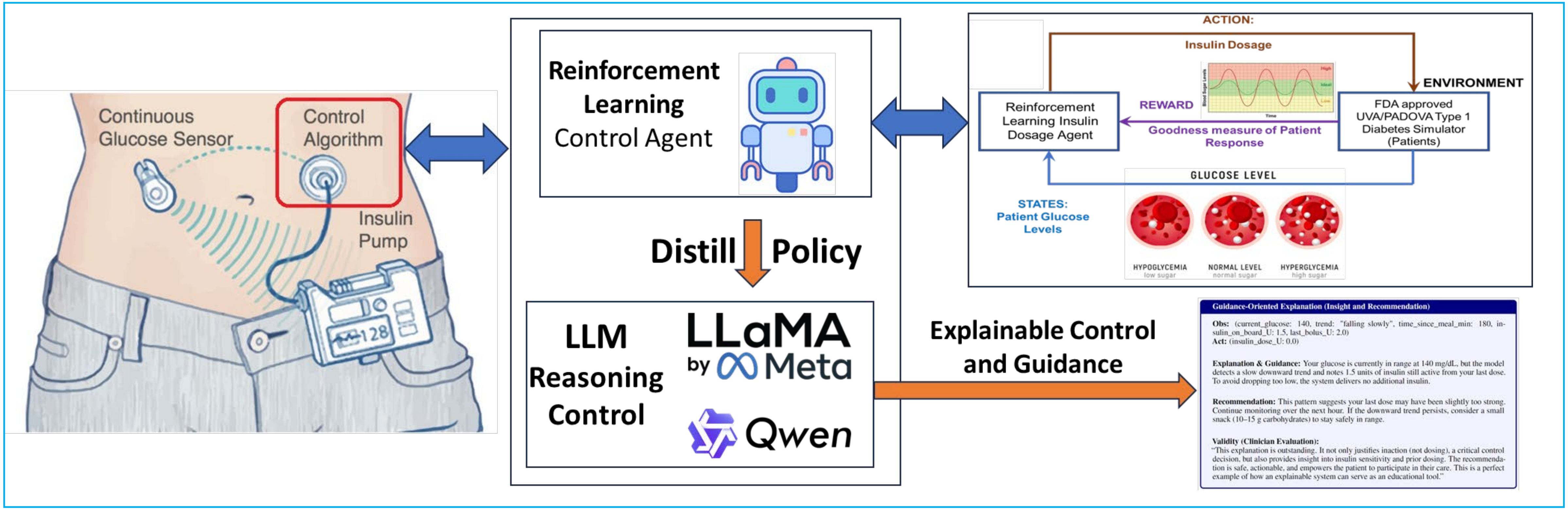


Image taken from iPAGScotland and modified

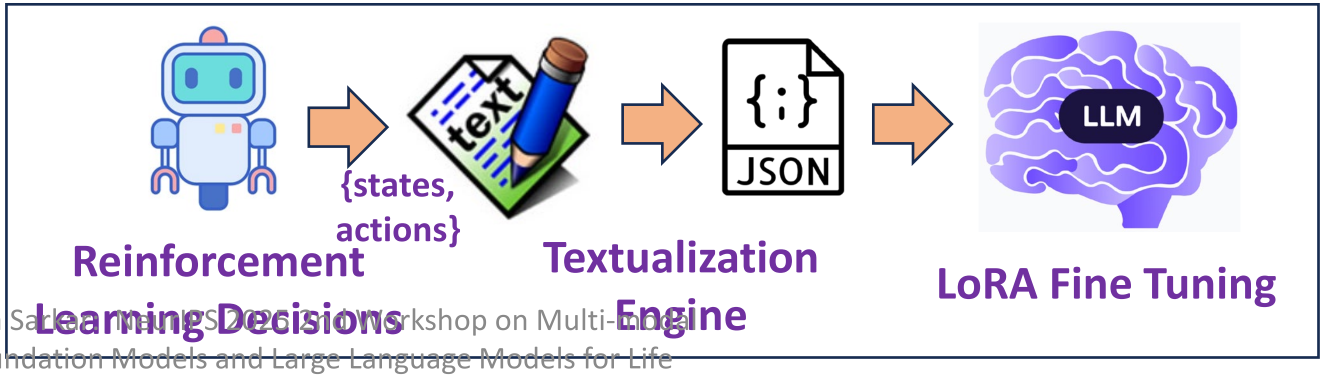


Reinforcement Learning Agent

Our Solution: Distilling Trust from Data

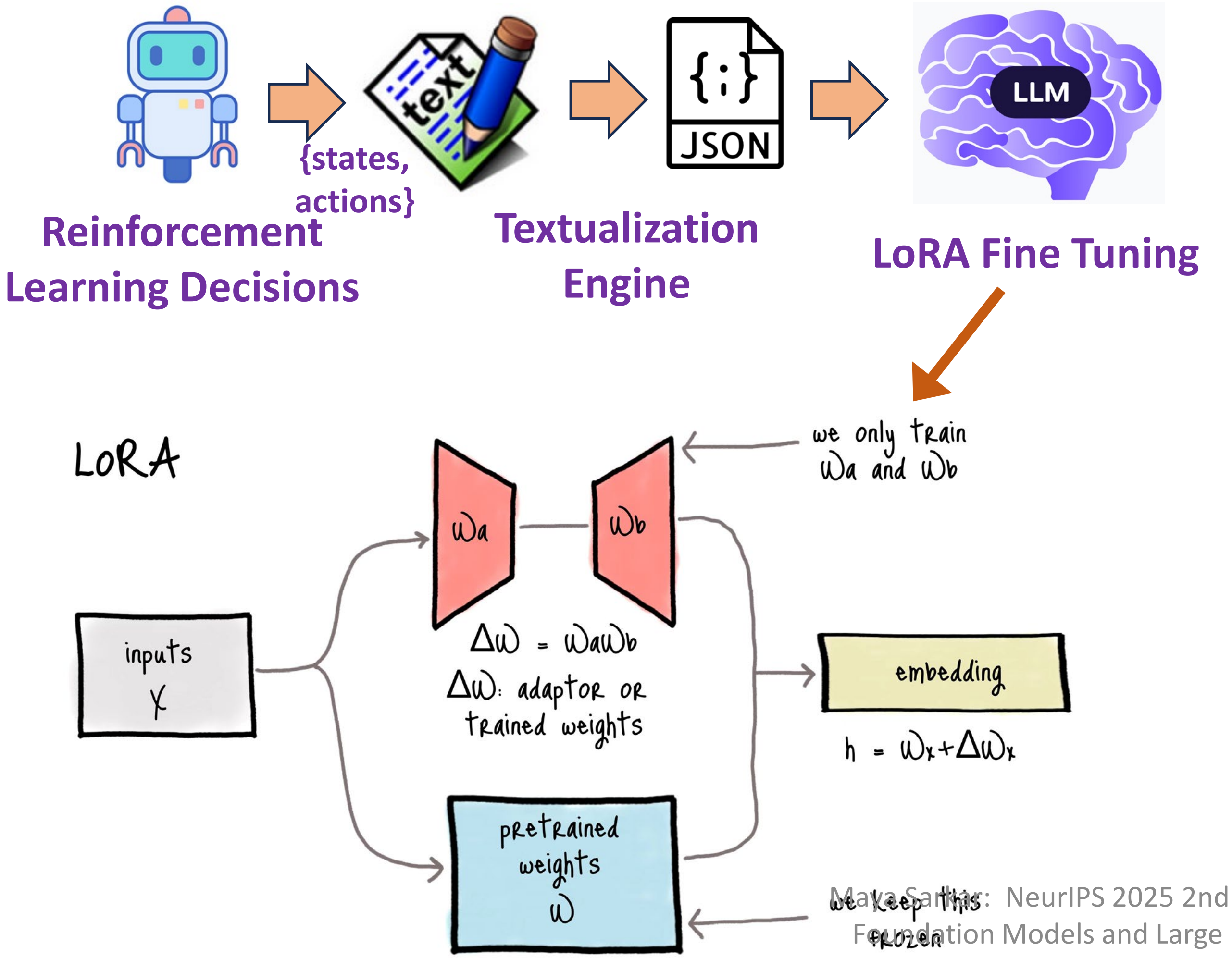


We teach a language model to think like an RL expert algorithm.



Our Solution: Distilling Trust from Data

We teach a language model to think like an expert algorithm.



1. **Train the Expert (RL Agent)**: Train a PPO RL agent in an FDA-approved T1D simulator, SimGlucose, to become an expert at controlling glucose levels automatically.
2. **Translate its "Thoughts" (Textualization Engine)**: We create a massive dataset of the RL agent's decisions. Our Textualization Engine converts these numerical decisions into structured, human-readable JSON prompts.
3. **Create the Explainer (LLM Controller)**: We fine-tune a Large Language Model (Llama 3.1 or Qwen3 8B) on this dataset. The LLM learns to make the same expert decisions, but because it's a language model, it can also *explain its reasoning*.

Explainability in Action

From Raw Data to Actionable Insight

The System Sees This (JSON):


```
{
  "current_glucose": "235 mg/dL",
  "trend": "rising slowly",
  "time_since_meal": "120 mins",
  ...
}
```


The Patient Sees This:

Explanation & Long-Term Guidance: Your glucose is high two hours after lunch. The system is delivering a **1.5 unit correction**.

Insight: We've noticed this happens often after your lunch. This may mean your standard lunch dose isn't quite enough.

Recommendation: At your next doctor's visit, you may want to **discuss adjusting your lunch-time insulin-to-carb ratio**.

LLM-Generated Decision & Explanation

Observation: Glucose is 210 mg/dL and rising 60 minutes after a meal. The last insulin dose was 2 hours ago.

Action: Deliver a correction bolus of 1.5 Units.

LLM Explanation: "Your glucose is currently high and still rising after your recent meal. To help bring it back into your target range, I am delivering a small correction dose of 1.5 units. This dose is calculated to be strong enough to lower your glucose without causing a rapid drop or 'crashing' later."

Validity: A human evaluation confirms this rationale is clinically sound, appropriate for the situation, and communicates the controller's intent effectively to the patient.

Short Explanation (Patient-Friendly Rationale)

Obs: (current_glucose: 195, trend: "rising", time_since_meal_min: 45, glucose_history_mg_dL: [180, 165, 150])
Act: (insulin_dose_U: 1.2)

Explanation: Your glucose is high and still rising after your recent meal. I'm giving you a small correction dose of 1.2 units to help guide your blood sugar back toward your target range.

Validity (Human Evaluation):
This explanation is clear, concise, and uses patient-friendly language. It correctly identifies the reason for the correction (post-meal hyperglycemia) and explains the action without causing alarm. It effectively communicates the system's intent.

Results: Performance with Trust

Achieving Top-Tier Control Without the Black Box

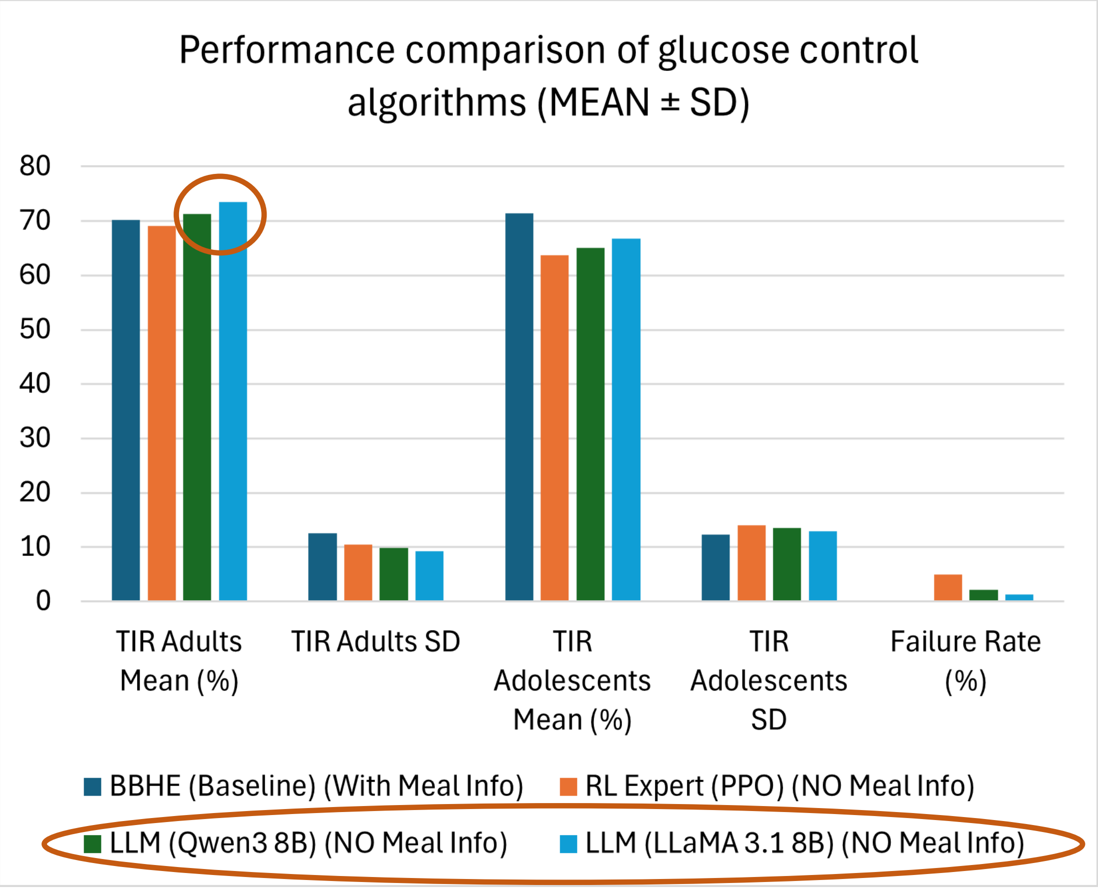
Quantitative Results: Time in Healthy Range (TIR)

•Our **RL→LLM Hybrid** controller achieves a higher Time in Range than the expert RL agent.

•It performs on par with clinical methods that require **full manual effort** (carb counting), while our system is **fully autonomous**.

Qualitative Results: Clinician-Verified Explanations

Clinician Validity: "This is an exemplary use of the technology... It transforms the device from a simple controller into a proactive health partner, facilitating more productive clinical conversations and leading to better long-term glycemic control."



Conclusion & Impact

A New Frontier for Trustworthy AI in Healthcare

•**We Bridged the Trust Gap:** Our framework successfully transforms an opaque "black box" controller into an explainable, trustworthy system.

•**We Empowered the Patient:** The system moves beyond simple automation to become an intelligent "copilot," providing insights that help patients understand their condition and collaborate more effectively with their doctors.

•**The Future is Explainable:** This distillation method provides a clear and impactful pathway for deploying advanced AI safely and effectively, not just in diabetes, but across the entire field of personalized medicine.

Guidance-Oriented Explanation (Insight and Recommendation)

Obs: (current_glucose: 140, trend: "falling slowly", time_since_meal_min: 180, insulin_on_board_U: 1.5, last_bolus_U: 2.0)
Act: (insulin_dose_U: 0.0)

Explanation & Guidance: Your glucose is currently in range at 140 mg/dL, but the model detects a slow downward trend and notes 1.5 units of insulin still active from your last dose. To avoid dropping too low, the system delivers no additional insulin.

Recommendation: This pattern suggests your last dose may have been slightly too strong. Continue monitoring over the next hour. If the downward trend persists, consider a small snack (10–15 g carbohydrates) to stay safely in range.

Validity (Human Evaluation):
This explanation is outstanding. It not only justifies inaction (not dosing), a critical control decision, but also provides insight into insulin sensitivity and prior dosing. The recommendation is safe, actionable, and empowers the patient to participate in their care. This is a perfect example of how an explainable system can serve as an educational tool.