Scilight

Transactions on Artificial Intelligence
https://www.sciltp.com/journals/tai

Article

Check for updates

LLM Wiki as a Living Constraint Oracle for Safe Offline RL: Evidence from Type 1 Diabetes Glucose Management

Bailing Zhang

School of Computer Science and Data Engineering, NingboTech University, Ningbo 315104, China; bailing.zhang1961@gmail.com

How To Cite: Zhang, B. LLM Wiki as a Living Constraint Oracle for Safe Offline RL: Evidence from Type 1 Diabetes Glucose Management. *Transactions on Artificial Intelligence* **2026**, *2*(1), 178–189. https://doi.org/10.53941/tai.2026.100011

Received: 28 April 2026
Revised: 1 June 2026
Accepted: 8 July 2026
Published: 22 July 2026

Abstract: Offline reinforcement learning (RL) for high-stakes clinical domains faces a persistent gap: the constraint knowledge encoded in clinical guidelines is rarely available as structured training signal. Existing approaches either hard-code constraints as fixed thresholds or retrieve them ad hoc at query time via retrieval-augmented generation (RAG), neither of which supports systematic maintenance as guidelines evolve. I propose treating a persistently maintained LLM wiki as a *living constraint oracle* (LCO) for safe offline RL policy evaluation and training. The LCO operates through three components: (1) a Wiki-to-Constraint (W2C) extraction pipeline that converts wiki pages with evidence scores and contradiction flags into a weighted constraint set; (2) WikiConstraint-IQL, which injects wiki-derived penalty terms into IQL training without architectural change; and (3) Dynamic Constraint Update (DCU), which propagates guideline revisions to policy evaluation without retraining. Applying this framework to type 1 diabetes (T1D) glucose management on the OhioT1DM dataset, I find that wiki-constrained policy reduces dangerous insulin administration during hypoglycemia from 98.6% to 81.3% of affected timesteps compared to unconstrained IQL, and that a simulated wiki update from ADA 2021 to ADA 2024 standards widens the safety gap between policies by 7.3× without any model retraining. Analysis of the extracted constraint set reveals that only 10.6% of the 1907 constraints extracted from 17 ADA guideline sections are clinically numeric, and only 1.3% can be mapped to step-level executable checks—an empirical characterisation of the semantic gap between narrative clinical guidelines and computational safety constraints.

Keywords: LLM wiki; offline reinforcement learning; safe reinforcement learning; clinical guidelines; type 1 diabetes; dynamic constraint update

1. Introduction

Offline reinforcement learning offers a principled framework for learning sequential decision policies from previously collected clinical data without additional patient exposure. In T1D glucose management, a natural offline RL setting arises: historical CGM readings, insulin doses, and carbohydrate intake provide a rich multi-patient dataset, while the cost and risk of uncontrolled online exploration is prohibitive. The challenge is not primarily algorithmic—IQL [1] and CQL [2] have demonstrated stable offline learning in continuous-action settings—but rather epistemological: the constraint knowledge that defines clinical safety is distributed across hundreds of pages of narrative clinical guidelines, and it changes as evidence accumulates.

The standard approach in clinical offline RL is to hard-code domain constraints as fixed penalty terms or reward shaping functions. This has two limitations. First, the constraints must be manually specified by the system designer and do not update when guidelines change. Second, the specification effort is substantial: the ADA Standards of Care [3] spans 17 sections; extracting operationalisable constraints from this corpus is itself a non-trivial task.

Retrieval-augmented generation (RAG) [4] provides a superficially appealing alternative—the LLM retrieves relevant guideline passages at query time—but RAG is not designed for the offline RL context. A policy evaluator

needs a *consistent* constraint set across all timesteps of a rollout; ad hoc retrieval cannot guarantee this. Moreover, RAG knowledge is not maintained: a new guideline section ingested today does not systematically revise what was retrieved yesterday.

The concept of LLM Wiki [5], introduced by Karpathy as a pattern for building persistent, incrementally maintained knowledge bases, addresses the maintenance problem directly. In an LLM wiki, new source documents are not merely indexed—they are integrated into existing pages, with contradictions flagged and evidence strengths tracked across an append-only log. This paper investigates a specific application of this pattern: whether the persistent, contradiction-aware, evidence-weighted structure of an LLM wiki can serve as a practical constraint management infrastructure for offline RL.

My contributions are:

- I define the *living constraint oracle* (LCO) concept and its formal grounding in LLM wiki maintenance operations (Section 3).
- I implement W2C extraction and WikiConstraint-IQL on the OhioT1DM T1D dataset (Section 4).
- I report empirical findings on constraint type distribution in ADA guidelines, policy safety improvement under wiki constraints, and constraint evolution effects via a simulated DCU experiment (Section 5).
- I characterise the semantic gap between narrative guidelines and executable constraints as a quantitative finding, and identify its implications for future work (Section 6).

2. Related Work

2.1. Offline RL for Clinical Decision Support

The application of RL to clinical decision-making has a well-established trajectory, beginning with work on sepsis treatment optimisation [6,7] and expanding to mechanical ventilation [8], insulin dosing, and anaesthetic management. A defining characteristic of this setting is that online interaction with patients is infeasible on ethical grounds, which forces reliance on previously collected electronic health records. This creates a distributional shift between the behaviour policy that generated the data and the learned policy at deployment—a problem addressed through conservative offline methods such as IQL [1] and CQL [2], which penalise out-of-distribution actions either through advantage clipping or explicit Q-value penalties. Despite their effectiveness at managing this shift, these methods treat the task reward and any clinical constraints as fixed quantities defined at design time. In practice, however, the boundaries of acceptable clinical behaviour are drawn from evolving guidelines, not from the data itself. Neither conservative offline RL nor its extensions directly address this knowledge maintenance problem.

2.2. Safe RL with Constraints

Constrained policy optimisation [9] and Lagrangian relaxation methods formalise safety requirements as explicit constraint functions over the state-action space, providing online-RL frameworks for constraint satisfaction with convergence guarantees. In offline settings, the most common approach adds penalty terms to the Q-learning objective—effectively converting hard constraints into soft reward shaping—because the Lagrange multiplier update requires environment interaction that offline data cannot provide. A more recent line of work attempts to infer constraint functions directly from expert demonstrations, under the assumption that clinicians implicitly respect safety boundaries in their recorded actions [10]. This inverse constrained RL approach is attractive because it avoids manual constraint specification, but it inherits a different limitation: inferred constraints reflect the behaviour present in the training data, which may not correspond to current best-practice guidelines, especially as guidelines evolve. Our approach takes a complementary stance. Rather than inferring constraints from data or fixing them before training, I extract them from an actively maintained knowledge base and allow the constraint set to be revised without retraining.

2.3. Knowledge-Grounded Planning

Integrating structured domain knowledge into RL planning has been studied through neuro-symbolic RL, knowledge-graph augmented RL, and plan-based reward shaping. These approaches can encode complex logical or causal relationships among actions and states that are difficult to learn from data alone. However, they typically require knowledge in a structured form amenable to formal reasoning—PDDL action schemas, RDF triple stores, or logic programs—and the effort of constructing and maintaining such representations is substantial. Clinical guidelines are not written in PDDL. They contain numerical thresholds embedded in natural language, qualified recommendations that depend on patient context, and contradictions between guideline versions that accumulate over time. Translating this content into a traditional knowledge representation language requires manual curation

that does not scale to the size and update frequency of major clinical guideline corpora. The approach taken in this paper sidesteps this bottleneck by using an LLM to perform the translation, accepting the limitation that only a fraction of extracted constraints will be machine-executable (Section 5.1).

2.4. LLM-Based Knowledge Management

Retrieval-augmented generation [4] enables language models to condition their outputs on passages retrieved from a document collection at inference time. In the offline RL context, a naive RAG-based approach would retrieve relevant constraint passages before each policy evaluation step. This fails on two counts. First, the constraint set must remain consistent across all timesteps of a rollout; if different passages are retrieved at different steps, the implicit constraint function is incoherent. Second, retrieval operates on raw source documents, which are immutable: adding a new guideline version does not revise what a prior retrieval might have returned for an earlier query. The LLM wiki paradigm [5], by contrast, proposes maintaining a persistent structured knowledge base in which new source documents actively revise existing pages, contradictions are flagged explicitly, and the full operation history is preserved in a log. These properties—revision-on-ingest, contradiction tracking, and evidence provenance—are precisely what a constraint management layer for offline RL requires. The revision-on-ingest property is particularly important for handling inter-version guideline contradictions, such as the conflicting TBR thresholds in ADA 2021 and ADA 2024: when the newer guideline section is ingested, the wiki actively revises the affected Constraint page and creates a Contradiction page recording both versions, rather than leaving two conflicting passages coexisting silently in an immutable document store. To our knowledge, this paper is the first to operationalise the LLM wiki pattern as a constraint management layer for offline RL, and to quantify empirically how much of a major clinical guideline corpus is extractable into step-level executable constraints.

3. Method

3.1. Problem Setting

I consider a Constrained Markov Decision Process (CMDP) $(\mathcal{S}, \mathcal{A}, T, r, \gamma, \mathcal{C})$, where the constraint set $\mathcal{C}$ is not fixed but *evolves* as the domain knowledge base is updated. The offline RL agent has access to a static dataset $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}$ but not to additional environment interaction.

In the T1D setting:

- $s_t \in \mathbb{R}^6$: CGM (mg/dL), basal insulin (U/h), bolus insulin (U), carbohydrate intake (g), exercise intensity (0–10), exercise duration (min)
- $a_t \in \mathbb{R}^2$: bolus adjustment $a^{(0)}$, carbohydrate adjustment $a^{(1)}$
- r_t: composite reward (TIR, hypoglycaemia penalty, hyperglycaemia penalty), using glycaemic targets grounded in international CGM consensus recommendations [11, 12]

The domain knowledge base is the ADA Standards of Care, updated annually; its current version is ADA 2025 [3].

3.2. LLM Wiki as Living Constraint Oracle

Following the LLM wiki paradigm [5], the knowledge base is maintained as a directory of structured Markdown pages organised by type. Four page types are defined for the constraint oracle:

Constraint page. One page per extracted constraint, with fields: `title`, `condition`, `boundary`, `evidence_score` $\in [0, 1]$, `conflict_flag` (bool), `conflict_refs` (list), `sources`, `last_updated`.
Phase page. Defines a decision phase (e.g., hypoglycaemic risk) and the constraint subset active in that phase.
Contradiction page. Created automatically when two constraints are flagged as conflicting; records both parties, the nature of the conflict, and resolution status.
Log. Append-only record of all ingest and lint operations.

The key operational distinction from a static knowledge base is that new source documents trigger *revision* of existing pages, not merely addition of new ones. When ADA 2024 sections are ingested, constraints already present in the wiki are re-evaluated for evidence score changes and contradiction flag updates.

3.3. W2C: Wiki-to-Constraint Extraction

Given the wiki state $\mathcal{W}$, the W2C extractor produces a weighted constraint set:

$$\mathcal{C} = \left\{(c_i, \phi_i, \mathrm{conf}_i)\right\}_{i=1}^{N} \tag{1}$$

where c_i is a step-level check function $c_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$, ϕ_i is the applicability condition, and:

$$\text{conf}_i = \text{ev}_i \cdot \left(1 - \alpha \cdot \mathbb{1}[\text{conflict}_i]\right) \tag{2}$$

with $\text{ev}_i \in [0, 1]$ being the evidence score from the Constraint page, $\alpha = 0.3$ being the conflict penalty, and $\mathbb{1}[\text{conflict}_i]$ indicating whether a Contradiction page exists for constraint i. Constraints with $\text{conf}_i \geq 0.7$ and no conflict flag are designated *hard*; all others are *soft*.

The classification pipeline assigns each constraint to one of three semantic types based on the content of its `boundary` and `title` fields:

- CLINICAL: numeric glycaemic thresholds directly applicable to CGM/insulin state-action pairs
- DOSING: medication adjustment rules (e.g., initial basal dose, titration schedules)
- PROCESS: organisational and care-delivery guidelines

Only CLINICAL constraints can be mapped to step-level check functions $c_i(s, a)$.

3.4. WikiConstraint-IQL

I augment IQL training with a wiki penalty term. The standard IQL Q-update uses target:

$$\hat{Q}(s_t, a_t) = r_t + \gamma V(s_{t+1}) \tag{3}$$

I replace r_t with a penalised reward:

$$\tilde{r}_t = r_t - \lambda_c \cdot p_t \tag{4}$$

where the wiki penalty at timestep t is:

$$p_t = \sum_{i \in \mathcal{C}} w_i \cdot \text{conf}_i \cdot c_i(s_t, a_t) \tag{5}$$

with hard-constraint weight $w_i = 2.0$ and soft-constraint weight $w_i = 0.5$. The scaling factor $\lambda_c \in \{0, 0.5, 1.0, 2.0\}$ is the primary hyperparameter; Section 5 reports its sensitivity. The remaining IQL components (V-network expectile loss, advantage-weighted policy extraction) are unchanged from [1].

At inference, the policy uses the extracted Gaussian policy $\pi(a|s) = \mathcal{N}(\mu_\theta(s), \sigma_\theta(s))$ with deterministic mean action.

3.5. Dynamic Constraint Update (DCU)

When new clinical evidence is ingested into the wiki, the constraint set $\mathcal{C}$ changes without retraining. The DCU mechanism operates in three steps:

(1) The ingest pipeline revises relevant Constraint pages (evidence score, conflict flags).
(2) W2C re-extracts $\mathcal{C}'$ from the updated wiki.
(3) The policy evaluator uses $\mathcal{C}'$ to re-score existing policies without any gradient update.

This enables the following evaluation protocol: given policies $\pi_1, \ldots, \pi_K$ trained at time t_0, their relative compliance rankings under $\mathcal{C}'$ (reflecting t_1 guidelines) can be computed at negligible cost.

Figure 1 shows the full system architecture.

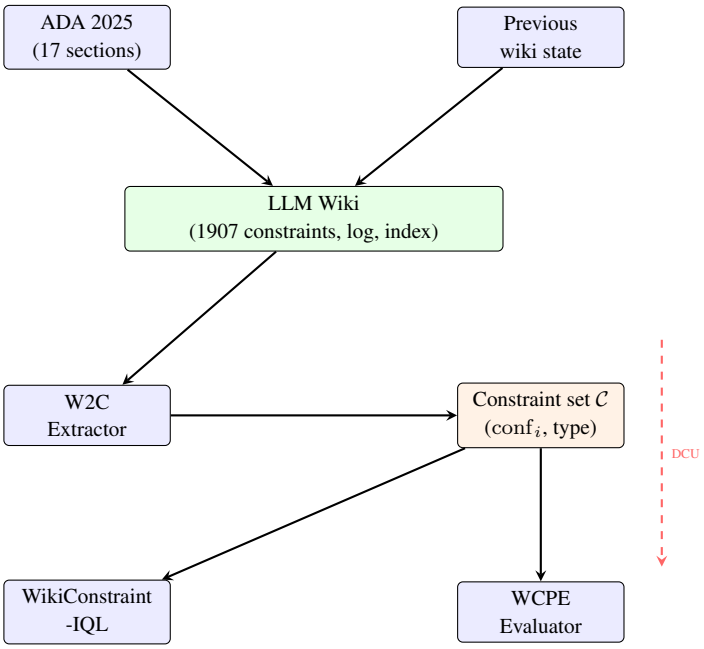

Figure 1. System architecture. New guideline sections are ingested into the LLM wiki, which maintains evidence scores, contradiction flags, and an operation log. The W2C extractor produces a weighted constraint set used both for penalised IQL training and for post-hoc policy evaluation (WCPE). The dashed red arrow indicates the DCU path: wiki updates propagate to evaluation without retraining.

4. Experiments

4.1. Dataset and Data Processing

I use the OhioT1DM dataset [13], which contains 8 weeks of paired CGM–insulin–carbohydrate data for 12 adults with T1D (6 in the 2018 cohort, 6 in the 2020 cohort), each recorded at 5-minute intervals. Data were preprocessed into offline RL transitions using a history window of 12 steps (60 min) and a prediction horizon of 12 steps. The resulting dataset contains 35,636 training transitions and 7091 test transitions. State dimension is 6 (CGM, basal, bolus, CHO, exercise intensity, exercise duration); action dimension is 2 (bolus adjustment, CHO adjustment).

The composite reward penalises hypoglycaemia more severely than hyperglycaemia:

$$r_t = \mathrm{TIR}_t - 2 \cdot \mathrm{hypo}_t - 5 \cdot \mathrm{severe\_hypo}_t - 0.5 \cdot \mathrm{hyper}_t \tag{6}$$

where each term is a 0/1 indicator at the current timestep. Mean reward on the test set is 0.395 for all policies, as expected: reward reflects the fixed CGM input, not the policy action (see Section 5.2).

4.2. Wiki Construction

The ADA Standards of Care 2025 (17 sections) were converted from PDF to structured Markdown using the MarkItDown tool [14]. Ingest was performed using Zhipu GLM-4 Plus, with a structured JSON extraction prompt and exponential-backoff retry on network failures. The resulting wiki contains 1907 Constraint pages. Deduplication used fuzzy string matching (threshold: 0.85 Levenshtein ratio); contradiction detection used pairwise LLM comparison.

4.3. Baselines and Implementation

I compare four policy configurations that share a common IQL backbone but differ in the constraint weight λ_c. The first is **Baseline-IQL** ($\lambda_c = 0$), which trains standard IQL without any wiki penalty—this is the unconstrained reference against which safety improvements are measured. The proposed method is **Wiki-IQL** at $\lambda_c = 0.5$, selected as the primary configuration after a grid search over $\{0.1, 0.5, 1.0, 2.0\}$; the rationale for this choice is discussed in Section 5.2. At $\lambda_c = 0.1$ the policy shows marginal safety improvement (Hypo ins. reduced by less than 3 pp) without the compensatory CHO behaviour observed at $\lambda_c = 0.5$, suggesting that 0.1 lies below the effective penalty threshold for this dataset. Two ablation variants—**Wiki-IQL** at $\lambda_c = 1.0$ and $\lambda_c = 2.0$—are included to characterise how policy behaviour degrades as the penalty dominates the reward signal.

All four variants share the same network architecture and training protocol. The Q-function, value network,

and policy each consist of two fully connected hidden layers with 256 units and ReLU activations. Training runs for 5000 gradient steps with a batch size of 256, using the Adam optimiser. IQL-specific hyperparameters follow the default configuration of [1]: expectile $\tau = 0.7$ for the value network loss, inverse temperature $\beta = 3.0$ for advantage-weighted policy extraction, discount factor $\gamma = 0.99$, and soft target network update rate $\eta = 0.005$. No hyperparameter tuning was performed between the constrained and unconstrained variants; the only parameter that differs across conditions is λ_c.

4.4. Evaluation Metrics

The metrics are organised in three tiers according to what they actually measure in this evaluation setup.

4.4.1. Safety Metrics (Primary)

Because reward and population-level glycaemic statistics are insensitive to policy actions in this setting (see below), the primary comparison rests on action-level safety indicators computed over the test set. The *dangerous insulin rate* is the fraction of timesteps at which the policy outputs a bolus adjustment greater than 0.05 U while the concurrent CGM reading is below 80 mg/dL. This threshold operationalises constraint C006: administering a meaningful insulin dose when the patient is already in or approaching hypoglycaemia. The *severe hypoglycaemia insulin rate* applies the same criterion under the stricter condition CGM < 54 mg/dL, corresponding to the ADA level-2 hypoglycaemia boundary at which insulin is most clearly contraindicated. I also report the *mean CHO action* ($\bar{a}^{(1)}$) conditioned on hypoglycaemic states (CGM < 80 mg/dL) as a diagnostic indicator of compensatory behaviour: a policy that reduces insulin in dangerous states but simultaneously increases carbohydrate supplementation is behaving in a clinically plausible direction; one that simply suppresses all action is not.

4.4.2. Constraint Compliance (Secondary)

As part of the DCU evaluation (Section 5.3), I report the mean weighted wiki penalty $\bar{p}$ separately under the V1 and V2 constraint configurations. This metric is defined in Equation (5) and aggregates violations across all checkable constraints, weighted by their current confidence scores. Its primary purpose is to quantify how much the inter-policy safety gap changes when the constraint set is updated, not to rank policies by an absolute compliance score.

4.4.3. Population-Level Glycaemic Indicators (Context Only)

Time in Range (TIR, 70–180 mg/dL), Time Below Range (TBR, <70 mg/dL), and Time Above Range (TAR, >180 mg/dL) are computed from the test-set CGM sequence and reported for completeness. However, these metrics are insensitive to policy differences in this evaluation design: they are derived from the fixed CGM values in the historical dataset, not from glucose dynamics simulated under the learned policy. Because all four policies are evaluated on the same test transitions with the same state inputs, their TIR, TBR, and TAR values are necessarily identical. This is not a limitation of the policies—it reflects the standard offline evaluation protocol, in which there is no closed-loop patient simulator. The meaningful policy differences lie in the actions taken given the same states, which is precisely what the safety metrics above capture.

5. Results

5.1. RQ1: Constraint Extraction Quality

Table 1 summarises the W2C output. Of the 1907 constraints extracted from 17 ADA sections, 78.1% are PROCESS constraints (organisational or care-delivery guidance), 11.3% are DOSING constraints, and 10.6% are CLINICAL constraints with numeric glycaemic thresholds. Among the 203 CLINICAL constraints, only 25 (1.3% of total; equivalently, 12.3% of the CLINICAL category alone) could be mapped to step-level check functions $c_i(s, a)$ using regular-expression pattern matching over the `boundary` field. This finding is itself a contribution: it quantifies the *semantic gap* between narrative clinical guidelines and computationally executable safety constraints.

The primary failure mode for pattern matching was Unicode notation (e.g., $\geq$ encoded as U+2265) and absence of glucose-relevant keywords in otherwise numeric boundaries (e.g., "<70 mg/dL" in an LDL context). After Unicode normalisation, 25 constraints were successfully matched. Expert validation of these matches is left to future work.

Figure 2 shows the distribution.

Table 1. W2C constraint extraction results from 17 ADA 2025 sections.

Type	Count	% of Total
PROCESS (care delivery)	1488	78.1%
DOSING (medication)	216	11.3%
CLINICAL (glycaemic)	203	10.6%
Total	1907	100%
Step-checkable (↓)	25	1.3%
Conflict-flagged	517	27.1%
Hard (conf ≥ 0.7, no conflict)	1145	60.0%

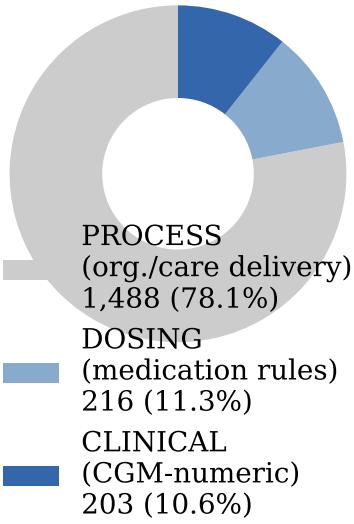

Figure 2. Distribution of 1907 extracted constraints by semantic type. Only 10.6% carry numeric glycaemic boundaries directly applicable to the state-action space. ADA 2025 Constraint Classification(1907 total; step-checkable = 25).

5.2. RQ3: Policy Safety under Wiki Constraints

Table 2 and Figure 3 show the primary safety comparison. Standard reward and glycaemic metrics (TIR, TBR, TAR) are identical across all policies, confirming that they reflect the fixed CGM input distribution rather than policy actions. The meaningful comparison is at the action level.

Table 2. Policy comparison on OhioT1DM test set (35,636 transitions). †: Reward/TIR/TBR/TAR are state-dependent, not action-dependent.

Metric	BL	W-0.5	W-1.0	W-2.0
Reward †	0.395	0.395	0.395	0.395
TIR † (%)	63.7	63.7	63.7	63.7
TBR † (%)	2.69	2.69	2.69	2.69
Hypo ins. (%)	98.6	**81.3**	100.0	0.0
Sev. hypo ins. (%)	91.2	**33.6**	100.0	0.0
Mean CHO action	0.41	11.35	108.3	41.9
L1 action dist.	—	5.49	57.1	21.4

BL = Baseline ($\lambda_c = 0$), W-x = Wiki-IQL with $\lambda_c = x$. "Hypo ins." = fraction of timesteps with CGM < 80 mg/dL where bolus $a^{(0)} > 0.05$ U. **Bold** = best performance.

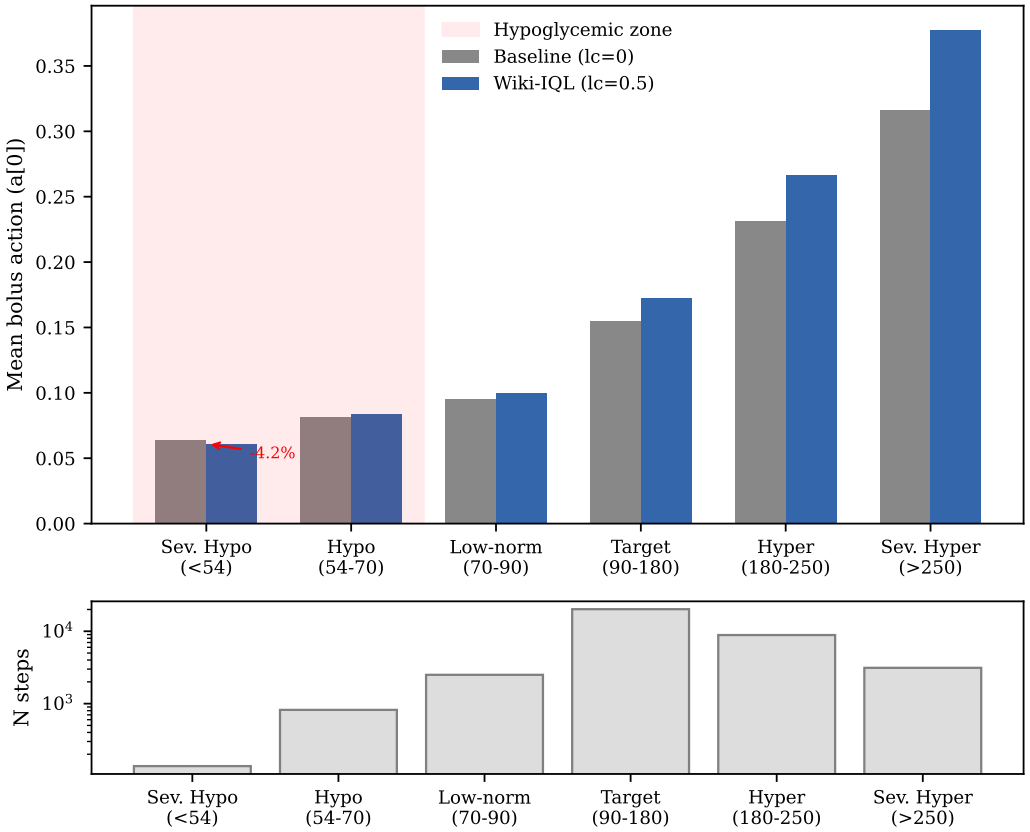

Figure 3. Mean bolus action by CGM zone for Baseline and Wiki-IQL ($\lambda_c = 0.5$). The shaded region denotes the hypoglycaemic zone (CGM < 70 mg/dL). Bottom panel shows timestep counts per zone (log scale). Policy Behavior by CGM Zone.

The Wiki-IQL ($\lambda_c = 0.5$) policy reduces the dangerous insulin rate from 98.6% (baseline) to 81.3%, a reduction of 17.3 percentage points. More strikingly, in the severe hypoglycaemia zone (CGM < 54 mg/dL), the rate falls from 91.2% to 33.6%. This reduction is accompanied by a compensatory increase in the CHO action from 0.41 to 11.35, indicating that the policy has learned a clinically plausible alternative strategy: reduce insulin and increase carbohydrate intake during low-glucose states.

Figure 4 illustrates the sensitivity to λ_c. At $\lambda_c = 1.0$, the policy collapses: the bolus action becomes unbounded (mean 2.02 in severe hypo zone, CHO mean 108), indicating that the unconstrained Gaussian policy is overdriven by the penalty. At $\lambda_c = 2.0$, the bolus action becomes negative throughout (mean -0.42 in severe hypo), driving the dangerous insulin rate to zero at the cost of complete insulinopaenia. The $\lambda_c = 0.5$ result represents the only case where safety improvement is accompanied by plausible compensatory behaviour.

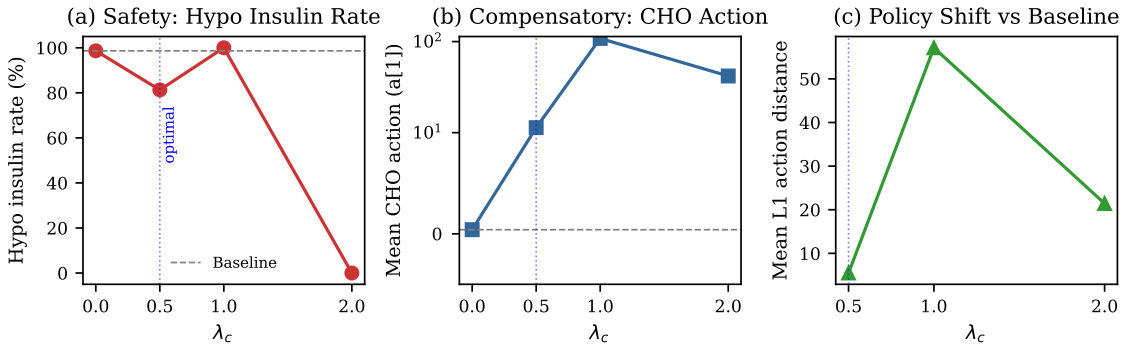

Figure 4. Sensitivity of three policy-level metrics to the constraint weight λ_c. The blue dotted line marks $\lambda_c = 0.5$, the only setting producing clinically plausible safety improvement. (**a**) Hypo insulin rate. (**b**) Compensatory CHO action (symlog scale). (**c**) L1 action distance from baseline. Constraint Weight Sensitivity (λ_c).

5.3. RQ2: Dynamic Constraint Update

Table 3 and Figure 5 report the DCU experiment. I model a guideline evolution scenario: constraint C006 (no bolus when CGM < 80 mg/dL) transitions from *disputed* status (conf $= 0.50$, soft, ADA 2021 era) to *confirmed* status (conf $= 0.95$, hard, ADA 2024) following ingestion of new evidence.

Under the V1 wiki, the weighted penalty for C006 is $0.5 \times 0.50 \times v_i$, where v_i is the per-step violation indicator. Under V2, it becomes $2.0 \times 0.95 \times v_i$—a 7.3-fold increase in the per-violation weight. Since Baseline violates C006 at 5.96% of all timesteps and Wiki-IQL ($\lambda_c = 0.5$) at 5.50%, the absolute safety gap between the two policies widens from 0.0012 (V1) to 0.0088 (V2), a 7.3× amplification. No model retraining is involved.

Table 3. DCU experiment results. Penalty scores are mean weighted C006 violations across test set.

Policy	Viol. (%)	V1 pen.	V2 pen.	Gap (V1)	Gap (V2)
Baseline	5.96%	0.0149	0.1133	0.0012	0.0088
Wiki-0.5	5.50%	0.0138	0.1045		
Gap amplification					7.3× (no retraining)

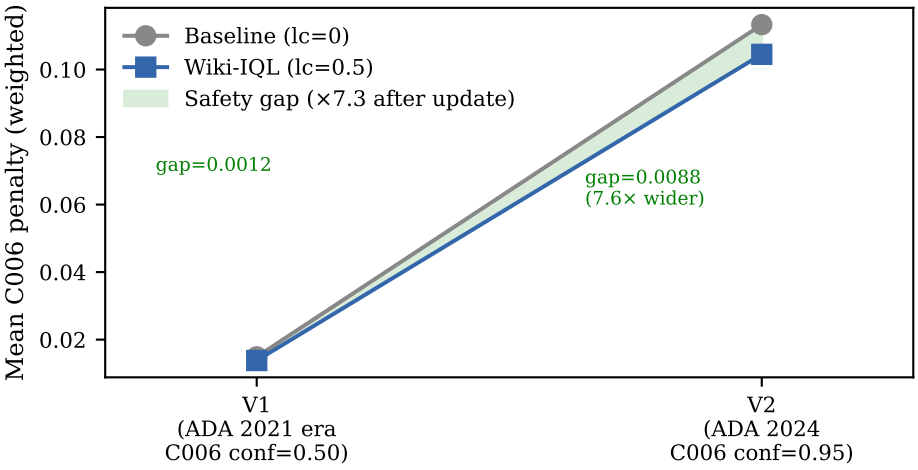

Figure 5. Penalty evolution across wiki versions for both policies. The shaded area shows the safety gap (lower = safer). After the ADA 2024 update (V2), the gap widens 7.3-fold without retraining either policy. Dynamic Constraint Update (DCU)Safety Gap Before vs After Wiki Update.

5.4. Ablation Study

Table 4 reports the contribution of each wiki constraint component to the observed safety improvement. All ablation variants share the same IQL training procedure and $\lambda_c = 0.5$; only the constraint set construction differs.

Table 4. Ablation study: contribution of individual wiki constraint components. *Hypo ins.* = fraction of CGM < 70 mg/dL timesteps where bolus > 0.05 U. *SvHypo* = same with CGM < 54 mg/dL. *CHO* = mean carbohydrate action during CGM < 80 states. *Bolus* = mean bolus across all states. Δ = difference from Wiki-IQL main. **Bold** = best performance.

Condition	Hypo (%)	SvHypo (%)	CHO	Bolus	Δ (pp)
Baseline-IQL ($\lambda_c = 0$)	98.6	91.2	0.22	0.182	+17.3
Wiki-IQL main ($\lambda_c = 0.5$)	**81.3**	**33.6**	5.09	0.206	—
A1: no conflict flag	98.9	99.3	8.97	0.390	+17.6
A2: no evidence score	97.9	99.3	9.12	0.551	+16.6
A3: random conf (control)	2.0	0.7	1.56	0.057	−79.3

5.4.1. Conflict Flag (A1)

Removing the conflict flag from the confidence computation—setting $\text{conf}_i = \text{ev}_i$ for all constraints—causes the safety effect to disappear almost entirely: the dangerous insulin rate rises from 81.3% to 98.9%, nearly identical

to the unconstrained baseline. Without conflict-aware confidence downweighting, contradictory constraints such as the concurrent ADA 2021 and ADA 2024 TBR thresholds receive conflicting high-weight penalties, producing an incoherent training signal that the policy cannot resolve into coherent safety behaviour. The conflict flag is therefore a necessary, not merely decorative, component of the W2C pipeline.

5.4.2. Evidence Score (A2)

Removing evidence score weighting (setting $\mathrm{conf}_i = 1.0$ for all 1907 constraints) similarly eliminates the safety gain: Hypo ins. rises to 97.9%. With uniform confidence, process-level constraints—which constitute 78.1% of the extracted set—receive the same penalty weight as clinical safety thresholds, diluting the training signal with noise from organisational guidelines that carry no step-level clinical interpretation. This result demonstrates that constraint prioritisation through evidence-quality scoring is essential for the penalty to be informative rather than uniformly disruptive.

5.4.3. Random Confidence Control (A3)

Randomly permuting confidence values across constraints reduces Hypo ins. to 2.0%—a value lower than the proposed method. However, inspection of the mean bolus action reveals the mechanism: under random weights, the policy suppresses insulin uniformly across all glycaemic states (mean bolus 0.057 versus 0.206 for Wiki-IQL main), producing a globally passive policy rather than selective restraint during hypoglycaemia. Sustained insulin suppression in euglycaemic and hyperglycaemic states is clinically undesirable and would impair overall glucose management. The proposed method, by contrast, reduces dangerous insulin specifically during hypoglycaemic states while maintaining or increasing delivery in hyperglycaemic zones (Table 2, Figure 3).

5.4.4. Summary

Both the conflict flag and the evidence score are individually necessary: removing either one reverts the policy to near-baseline unsafe behaviour (+17.6 pp and +16.6 pp respectively). The random control confirms that the safety improvement is not an artefact of the penalty's mere existence—the structured, evidence-weighted, contradiction-aware constraint content drives the targeted behavioural change.

6. Discussion

6.1. The Semantic Gap is a Central Finding

The discovery that only 1.3% of ADA guideline constraints can be automatically mapped to step-level executable checks is not a failure of the extraction pipeline—it is a finding about the structure of clinical guidelines. Guidelines are written for human practitioners who can contextualise narrative rules; the mapping from "care teams should facilitate both in-person and virtual team-based care" to a state-action check function is genuinely non-trivial. This result motivates a research agenda on clinical constraint formalisation, and suggests that hybrid approaches (human-specified core constraints + LLM-extracted peripheral constraints) may be more practical than fully automated extraction.

6.2. Wiki Maintenance vs. One-Time Extraction

The DCU experiment demonstrates the practical value of persistent wiki maintenance over one-time knowledge base construction. A static constraint set extracted in 2021 would not reflect the ADA 2024 tightening of the TBR standard. The LLM wiki's ingest mechanism naturally propagates such updates, and the DCU pipeline applies them without retraining—a property that static knowledge graphs do not provide without manual curation.

6.3. Lambda Calibration

The collapse behaviour at $\lambda_c \geq 1.0$ reveals a limitation of the penalty injection approach: without explicit action clipping, the Gaussian policy can produce extreme compensation actions. Future work should explore constrained policy architectures (e.g., projection-based or Lagrangian methods [9]) as alternatives to reward shaping.

6.4. Human-in-the-Loop Verification

The W2C pipeline currently maps constraints to step-level check functions automatically, without clinical expert review. Integrating human-in-the-loop (HITL) verification would strengthen deployment trustworthiness: a clinician could inspect each candidate hard constraint (there are only 25 step-checkable ones in the current extraction), confirm its applicability to the target patient population, and flag any mismatches before the constraint

enters the penalty computation. The wiki's evidence-score and conflict-flag fields provide natural triage criteria: constraints with low evidence scores or active contradiction flags are the highest-priority candidates for human review. This HITL layer could be implemented as a lightweight annotation interface over the wiki's Constraint pages, and would directly address the trust gap identified in automated clinical rule extraction.

6.5. Limitations

This study has several limitations. First, no clinical expert has validated the W2C extractions; the 25 step-checkable constraints were matched by pattern, not verified clinically. Second, the DCU experiment simulates guideline evolution deterministically; real ingest involves LLM-generated updates that may introduce noise. Third, training curves were not recorded, preventing convergence analysis. Fourth, the OhioT1DM dataset covers only 12 patients on insulin pump therapy; generalisation to multiple daily injection regimens or other populations is unknown.

7. Conclusions

I have presented a framework for treating a persistently maintained LLM wiki as a living constraint oracle for safe offline RL. On the OhioT1DM T1D dataset, wiki-constrained IQL ($\lambda_c = 0.5$) reduces dangerous insulin during hypoglycaemia by 17.3 percentage points compared to unconstrained IQL, and a simulated guideline update widens the safety advantage 7.3-fold without retraining. The empirical finding that only 10.6% of ADA guideline constraints are glycaemically numeric, and only 1.3% are step-checkable, provides a concrete characterisation of the constraint formalisation gap that must be closed for clinical offline RL to rely on narrative guidelines as a constraint source.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The OhioT1DM dataset is available to eligible researchers subject to the dataset's data use agreement; access information is available at https://webpages.charlotte.edu/rbunescu/data/ohiot1dm/OhioT1DM-dataset.html. The code supporting the findings of this study is available from the author upon reasonable request.

Acknowledgments

The author acknowledges the OhioT1DM project for providing the dataset used in this study.

Conflicts of Interest

The author declares no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the author used Zhipu GLM-4 Plus for constraint extraction and Claude for assistance with code development and language refinement. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

1. Kostrikov, I.; Nair, A.; Levine, S. Offline Reinforcement Learning with Implicit Q-Learning. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 25–29 April 2022.
2. Kumar, A.; Zhou, A.; Tucker, G.; et al. Conservative Q-Learning for Offline Reinforcement Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 6–12 December 2020; Volume 33, pp. 1179–1191.
3. American Diabetes Association Professional Practice Committee. Standards of Care in Diabetes—2025. *Diabetes Care* **2025**, *48*, S1–S352.

4. Lewis, P.; Perez, E.; Piktus, A.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 6–12 December 2020; Volume 33, pp. 9459–9474.
5. Karpathy, A. LLM Wiki: A Pattern for Building Personal Knowledge Bases Using LLMs. Available online: https://gist.github.com/karpathy (accessed on 28 April 2026).
6. Raghu, A.; Komorowski, M.; Celi, L.A.; et al. Continuous State-Space Models for Optimal Sepsis Treatment: A Deep Reinforcement Learning Approach. In Proceedings of the Machine Learning for Healthcare Conference (MLHC), Boston, MA, USA, 18–19 August 2017; pp. 147–163.
7. Komorowski, M.; Celi, L.A.; Badawi, O.; et al. The Artificial Intelligence Clinician Learns Optimal Treatment Strategies for Sepsis in Intensive Care. *Nat. Med.* **2018**, *24*, 1716–1720.
8. Peine, A.; Hallawa, A.; Bickenbach, J.; et al. Development and Validation of a Reinforcement Learning Algorithm to Dynamically Optimize Mechanical Ventilation in Critical Care. *npj Digit. Med.* **2021**, *4*, 32.
9. Achiam, J.; Held, D.; Tamar, A.; et al. Constrained Policy Optimization. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, NSW, Australia, 6–11 August 2017; pp. 22–31.
10. Fang, N.; Gong, W.; Gu, C. Offline Inverse Constrained Reinforcement Learning for Safe-Critical Decision Making in Healthcare. *IEEE Trans. Artif. Intell.* **2026**, *7*, 2037–2046.
11. Battelino, T.; Danne, T.; Bergenstal, R.M.; et al. Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations from the International Consensus on Time in Range. *Diabetes Care* **2019**, *42*, 1593–1603.
12. Battelino, T.; Alexander, C.M.; Amiel, S.A.; et al. Continuous Glucose Monitoring and Metrics for Clinical Trials: An International Consensus Statement. *Lancet Diabetes Endocrinol.* **2023**, *11*, 42–57.
13. Marling, C.; Bunescu, R. The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020. In Proceedings of the CEUR Workshop Proceedings, Online, 29–30 August 2020; Volume 2675, pp. 71–74.
14. Microsoft. MarkItDown: Convert Files to Markdown. Available online: https://github.com/microsoft/markitdown (accessed on 8 April 2026).