Wiki-Constrained Offline Reinforcement Learning for Strategy Recommendation in Health Support Dialogue: A Proof-of-Concept Study

Bailing Zhang
School of Computer Science and Data Engineering,
NingboTech University, No. 1 Qianhu South Road,
Ningbo, Zhejiang 315100, China

Ruofan Zhang
School of Psychology,
The University of Queensland,
Brisbane, QLD 4072, Australia

ABSTRACT

Aim: To develop and evaluate a proof-of-concept framework, Wiki-Constrained Conservative Hyperbolic Implicit Q-Learning (Wiki-CH-IQL), that integrates clinical strategy constraints from a persistently maintained Therapeutic Constraint Wiki (TC-Wiki) into an offline reinforcement learning (RL) system for strategy recommendation in health support dialogues, with particular focus on the Dynamic Constraint Update (DCU) mechanism that enables guideline revision without model retraining.

Materials and Methods: TC-Wiki stores stage-binding, sequence, contraindication, and frequency constraints as structured pages with evidence scores and conflict flags, extracted from clinical counselling texts using a large-language-model pipeline (W2C). Wiki-derived penalty terms modify the offline RL reward without architectural change. Experiments used the ESConv emotional support dialogue dataset (1,300 dialogues; 11,899 training, 2,518 test transitions). A controlled ablation compared the full system (penalty weight $\lambda c = 0.5$) against the unconstrained baseline ($\lambda c = 0$). A DCU experiment simulated guideline revision by upgrading a single constraint from disputed to confirmed status.

Results: Wiki-CH-IQL ($\lambda c = 0.5$) achieved a 24.48% match rate with expert behaviour and 16.47% macro recall. The ablation ($\lambda c = 0$) yielded 24.32% and 16.23%, confirming that wiki penalties contribute consistently but marginally (+0.16 pp match rate, +0.24 pp macro recall). Strategy shifts toward questioning (+49.9 pp) and suggestions (+11.7 pp) were consistent with Helping Skills Theory. The DCU experiment demonstrated a 2.3× amplification of the safety gap without retraining. Policy concentration on two strategies was present in both conditions, confirming collapse as a base RL training issue.

Conclusions: The proof-of-concept demonstrates that clinical counselling guidance can be partially formalised as an updatable constraint wiki and integrated into offline RL. The DCU mechanism is the primary practical contribution, enabling knowledge base revisions to propagate to policy evaluation without retraining. Limitations include simulated constraint sources and policy collapse from the base RL training. These define concrete next steps toward clinician-validated health dialogue decision support systems.

General Terms

Algorithms; Health Informatics; Machine Learning; Reinforcement Learning; Decision Support Systems.

Keywords

Health informatics; offline reinforcement learning; decision support systems; knowledge base; strategy recommendation; dynamic constraint update.

1. INTRODUCTION

Health dialogue support systems face a challenge that statistical learning approaches handle poorly: clinical guidelines regulating appropriate strategy sequencing, contraindications under emotional distress, and limits on repetitive interventions cannot be inferred from annotated corpora alone. These constraints are formalised in counselling psychology frameworks such as Hill's Helping Skills Theory [1], which structures support interactions into Exploration, Insight, and Action phases. Automated chatbots, including cognitive behavioural therapy-grounded systems [2], have demonstrated preliminary feasibility but do not natively enforce theoretical sequencing. Grodniewicz and Hohol [3] identify this absence as a central barrier to AI-assisted psychotherapy.

Offline reinforcement learning (RL) addresses policy learning from fixed demonstration data without live client interaction, which is ethically appropriate for clinical contexts [4,5,6]. Conservative Hyperbolic Implicit Q-Learning (CH-IQL) [7] frames strategy selection in emotional support dialogues as an offline MDP, combining Implicit Q-Learning [5] with Conservative Q-Learning regularisation [6] and a hierarchy-aware Poincaré ball embedding. The resulting policy shifts toward question-asking and suggestion-giving, consistent with counselling theory. However, the training process does not consult clinical guidelines: the policy is constrained only by data distribution, not by validated clinical principles.

A complementary development is the living constraint oracle concept [8], which treats a persistently maintained large-language-model (LLM) wiki as a structured constraint management layer for offline RL, demonstrating that guideline updates propagate to policy evaluation without model retraining. The present paper applies this concept to health support dialogue, where constraints are qualitative rather than numeric.

The central contribution is the Therapeutic Constraint Wiki (TC-Wiki) framework and its Dynamic Constraint Update (DCU) mechanism. Rather than claiming that constraint injection substantially improves policy metrics—a claim that the ablation study reported here shows to be marginal at the current hyperparameter setting—this paper focuses on the architectural property that enables knowledge revision without retraining: a practically important requirement in real-world health AI where clinical guidelines evolve continuously.

The specific contributions are: (i) a four-type therapeutic constraint taxonomy (stage-binding, sequence, contraindication, frequency) with a wiki page schema; (ii) a Wiki-to-Constraint (W2C) LLM extraction pipeline and characterisation of the semantic gap between narrative clinical guidance and executable checks; (iii) Wiki-CH-IQL, which injects wiki penalties into

offline RL without architectural change; and (iv) a DCU experiment quantifying how guideline revision propagates to policy evaluation.

This study is a proof of concept. No clinical validation with real help-seekers was performed. The system is a research prototype for clinician-supervised decision support, not autonomous deployment.

2. MATERIALS AND METHODS

2.1 Therapeutic Constraint Wiki (TC-Wiki)

TC-Wiki organises clinical strategy guidance into five constraint types, four of which are computationally executable (Table 1). Stage-binding constraints specify which strategies are appropriate in the Exploration, Insight, or Action phase. Sequence constraints require that certain strategies be preceded by others within a rolling history window of K turns. Contraindication constraints prohibit strategies when estimated emotional intensity exceeds a threshold. Frequency constraints limit consecutive or cumulative repetition. Process constraints capture relational guidance that does not map to step-level checks and are excluded from penalty computation.

Table 1. Constraint taxonomy used in TC-Wiki. State requirement column lists extended state components read by check function ci.

Type	Example rule	State requirement	Executable
Stage-binding	No suggestions in Exploration without Reflection	stage, action	Yes
Sequence	Suggestions require Reflection in history window	history, action	Yes
Contraindication	No challenge when emotional intensity > 0.85	emotion level, action	Yes
Frequency	Same strategy ≤ 3 consecutive turns	history, counts	Yes
Process	Facilitate rapport through active listening	—	No

2.2 W2C Extraction Pipeline

The Wiki-to-Constraint (W2C) pipeline converts clinical text into the TC-Wiki constraint set in three stages: (i) section-aware chunking (500-word segments with 50-word overlap); (ii) structured JSON extraction using GLM-4 (Zhipu AI, Beijing, China) with a constrained prompt that classifies each extracted constraint by type and produces a check-logic sketch; (iii) deduplication, ID assignment, and pairwise contradiction detection for constraints targeting the same strategy with opposing action constraints.

Because the published clinical manuals (Hill, 2014 [1]; Beck et al., 1979 [9]; Miller and Rollnick, 2013 [10]) were not available in machine-readable form at the time of these experiments, the pipeline was applied to three author-constructed simulated texts totalling 2,566 words, written to reproduce the genre and constraint structure of counselling psychology manuals. This is a significant limitation: the high executable rate observed (88.5%) reflects the constraint-oriented construction of these texts rather than the actual yield of real published sources. All claims about the TC-Wiki constraint yield require replication using the actual published materials.

The pipeline extracted 52 constraints, of which 46 (88.5%) are executable as step-level check functions. All non-executable constraints are of type Process. Table 2 summarises the extracted constraint set.

Table 2. TC-Wiki constraint extraction results from three simulated source texts (2,566 words total). The high executable rate reflects constraint-oriented text construction; rates from actual published manuals are expected to be substantially lower.

Type	Count	Executable	Hard	Soft
Stage-binding	12	12	4	8
Sequence	11	11	3	8
Contraindication	9	9	2	7
Frequency	14	14	0	14
Process	3	0	0	0
Total	**52**	**46**	**9**	**37**

2.3 Wiki-CH-IQL

Wiki-CH-IQL extends CH-IQL [7] with a wiki-derived penalty layer. The extended state representation $s_t^{ext} \in R^{117}$ concatenates: the Poincaré ball encoding of the supporter utterance $h_t \in R^{64}$; a one-hot therapeutic stage indicator (E/I/A); a K = 5-turn strategy history encoding (40 dimensions); estimated emotion intensity $emo_t \in [0,1]$; normalised per-strategy usage counts (8 dimensions); and normalised turn index. The extended state provides all information required for constraint evaluation.

For each training transition (s_t, a_t), the wiki penalty is computed as: $p_t = \Sigma\, w_i \times conf_i \times c_i(s_t^{ext}, a_t)$, where ci is the step-level check function for constraint i. The penalty modifies the training reward: $r_tilde_t = r_t - \lambda c \times p_t$, with $\lambda c = 0.5$. This penalised reward replaces the raw reward in the Q-network update target without any change to the network architecture.

The full training objective combines the modified Q-loss (dual networks with temporal-difference target $r_tilde_t + \gamma V(s_{t+1})$), expectile regression for the value network ($\tau = 0.6$), advantage-weighted policy extraction with behavioural cloning ($\lambda BC = 0.3$), log-sum-exp conservative regularisation ($\lambda CQL = 0.5$), and Poincaré hierarchy margin loss ($\lambda hier = 0.2$). The constraint penalty is the only modification relative to the original CH-IQL algorithm [7].

The Dynamic Constraint Update (DCU) mechanism allows constraint confidence scores to be revised without retraining either policy. Given two wiki versions V1 and V2 differing only in the confidence of constraint cj, the amplification ratio ρ = [ρp^pi0(V2) − ρp^pi1(V2)] / [ρp^pi0(V1) − ρp^pi1(V1)] quantifies how much the safety gap between constrained and unconstrained policies changes following the wiki revision. This provides a practical mechanism for propagating guideline updates to deployed policies. Figure 1 shows the overall system architecture.

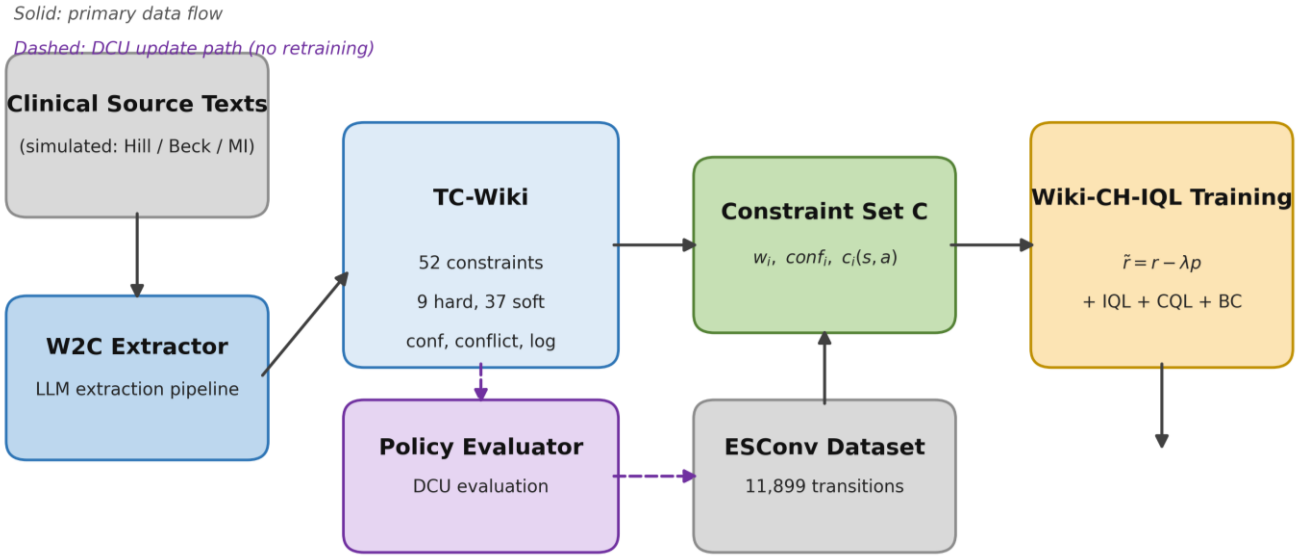

Figure 1. Wiki-CH-IQL system architecture. Clinical source texts are processed by the W2C extractor to populate TC-Wiki with typed constraint pages. The dashed arrow indicates the DCU path: wiki updates propagate to policy evaluation without retraining. The penalty pt modifies the offline reward before Q-network updates

2.4 Dataset and Training Setup

Experiments used the ESConv dataset [11], comprising 1,300 multi-turn dialogues between trained supporters and help-seekers experiencing emotional difficulties. Strategy labels are stored in the annotation field of each supporter turn using natural-language names; case-insensitive normalisation recovered 16,940 labelled turns including 1,302 previously lost due to preprocessing case mismatch. Dialogues were split 70/15/15 by dialogue into 11,899 training, 2,523 validation, and 2,518 test transitions.

A stage-aligned reward (rt ∈ {−1, 0, +1}) was used: +1 if action at is preferred for the detected therapeutic stage, −1 if penalised, 0 otherwise. This differs in scale from the advantage-based reward in the original CH-IQL [7]; absolute policy improvement and balanced score values are therefore not cross-system comparable. The text encoder is a hash-based bag-of-words projection (MLP 256→56→64), replacing the pre-trained sentence encoder of the original implementation to avoid external dependencies. All other network components are identical to CH-IQL [7]. Training used: learning rate 10^-4, batch size 64, maximum 150 epochs with early stopping (patience 30), gradient norm clipping 0.5, Q-value clamping [−50, 50].

Ethics statement: This study is a computational experiment on a publicly available anonymised corpus. No human participants were recruited and no personal data were collected or processed. The ESConv dataset was collected under informed consent procedures described in the original publication [11].

3. RESULTS

3.1 Match Rate, Macro Recall, and Ablation

Table 3 reports match rate, macro recall, and policy improvement for the behavior policy baseline, the original CH-IQL system [7], the unconstrained ablation (λc = 0), and the full Wiki-CH-IQL (λc = 0.5). The ablation (λc = 0) and full model use identical architectures and reward functions, enabling direct controlled comparison.

Table 3. Main results and ablation on the ESConv test set. CH-IQL numbers are from the published paper [7]. Policy improvement uses a different reward scale from CH-IQL and is not cross-system comparable; match rate and macro recall share the same definition. Dagger: reward scales differ; see Section 2.4.

Method	Match (%)	Improvement	Macro Recall (%)
Behavior policy	100.0	0.000	—
CH-IQL [7]	26.32	+0.188	16.79
Wiki-CH-IQL λc = 0 (ablation)	24.32	1.014	16.23
Wiki-CH-IQL λc = 0.5 (full)	**24.48**	**1.059**	**16.47**

Wiki-CH-IQL (λc = 0.5) achieves 24.48% match rate and 16.47% macro recall. The ablation (λc = 0) yields 24.32% and 16.23%. The wiki penalty contributes +0.16 percentage points in match rate and +0.24 pp in macro recall. These margins are small but directionally consistent across both metrics, confirming a positive contribution of the constraint layer at this hyperparameter setting.

The modest effect size is technically expected. Of the 46 executable constraints, 37 are soft (wi = 0.5) and 9 are hard (wi

= 2.0). With λc = 0.5, the maximum per-step penalty from a single hard constraint at maximum confidence (conf = 0.90) is 2.0 × 0.90 × 0.5 = 0.90, comparable in magnitude to the stage-aligned reward signal (±1.0). Constraint violations are also sparse in practice, as not all constraints fire at every turn. A larger λc would amplify the wiki effect but at the cost of training stability [8]. The current result establishes the mechanism as operational; systematic sensitivity analysis is deferred to future work.

Expressed in relative terms, the wiki penalty improves match rate by 0.66% (0.16 / 24.32) and macro recall by 1.48% (0.24 / 16.23) over the ablation, a consistent positive shift across both metrics. Relative to the original CH-IQL system [7], the full model attains a lower absolute match rate (24.48% versus 26.32%, a difference of 1.84 pp) and macro recall (16.47% versus 16.79%). This gap is attributable to the two deliberate simplifications documented in Section 2.4—the stage-aligned reward on a different scale and the hash-based bag-of-words encoder that replaces the pre-trained sentence encoder—rather than to the constraint layer, because the ablation confirms that removing the wiki penalty leaves both metrics essentially unchanged. Table 4 consolidates the per-metric and per-strategy contribution of the wiki penalty isolated by the controlled ablation, and Figure 4 shows the same quantities graphically.

Table 4. Contribution of the wiki penalty isolated by the controlled ablation (λc = 0 → λc = 0.5) on the ESConv test set. Metric values are taken from Table 3 and strategy values from Table 5. Relative change is the absolute change divided by the λc = 0 value. All shifts are small but directionally consistent.

Quantity	λc = 0 (ablation)	λc = 0.5 (full)	Absolute change	Relative change
Match rate (%)	24.32	24.48	+0.16 pp	+0.66%
Macro recall (%)	16.23	16.47	+0.24 pp	+1.48%
Question (%)	70.2	69.0	−1.2 pp	−1.71%
Providing Suggestions (%)	26.3	27.0	+0.7 pp	+2.66%
Information (%)	3.4	4.0	+0.6 pp	+17.6%

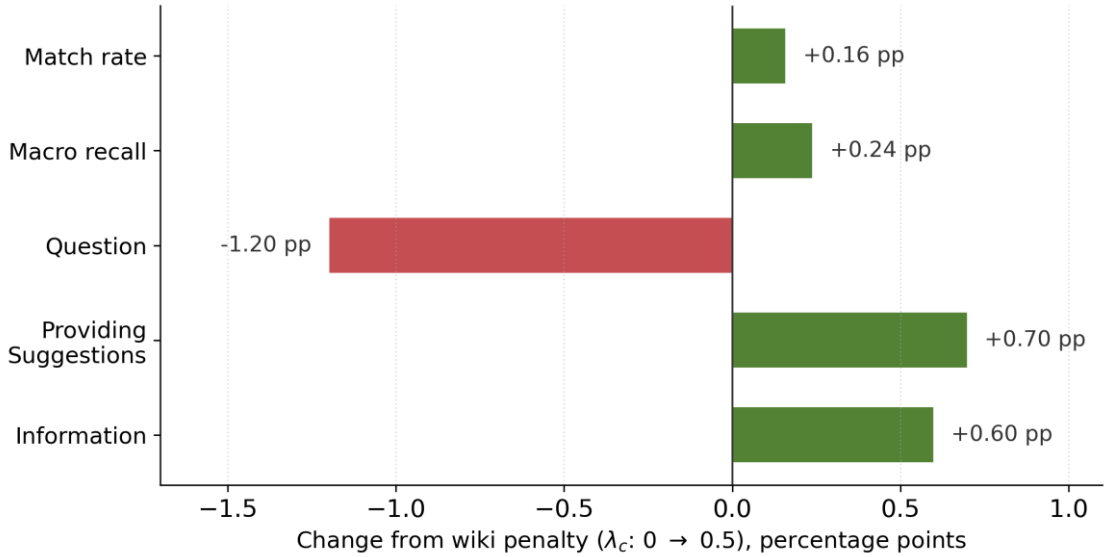

Figure 4. Wiki-penalty contribution isolated by the controlled ablation (λc = 0 → λc = 0.5). Bars show the absolute change in percentage points for the two headline metrics and the three strategies whose frequency changes; positive values (green) indicate an increase under the wiki penalty and negative values (red) a decrease. The magnitudes confirm a small but consistent corrective effect.

3.2 Strategy Distribution

Figure 2 and Table 5 show strategy frequencies for the behavior policy, the unconstrained ablation (λc = 0), and the full Wiki-CH-IQL (λc = 0.5). The two policy conditions are strikingly similar: both concentrate outputs on Question (~70%) and Providing Suggestions (~27%), while five of eight strategies receive zero probability mass. Wiki penalties marginally reduce Question frequency (−1.2 pp) and increase Suggestions (+0.7 pp) relative to the ablation.

The critical finding is that policy collapse is present in both conditions. This confirms that the collapse is a property of the base offline RL training—driven by sparse reward, small training set (909 dialogues), and the simplified text encoder—rather than an artefact of the wiki constraint injection. From a health informatics safety perspective, a decision support system that never employs reflective listening, affirmation, or restatement is unsafe for any user-facing application, regardless of how it is constrained. Addressing this collapse through reward redesign and richer state encoding is the most urgent direction for future work.

Quantitatively, the two learned policies place 96.5% (ablation) and 96.0% (full model) of their probability mass on Question and Providing Suggestions combined, compared with 34.4% for the behavior policy. Five strategies—Reflection of Feelings, Restatement or Paraphrasing, Self-disclosure, Affirmation and Reassurance, and Others—receive exactly zero mass under both policy conditions, although together they account for 59.8% of the behavior-policy distribution (7.7%, 7.7%, 10.5%, 15.7%, and 18.2% respectively). Information is the only minority strategy

the policies retain, at 3.4–4.0% versus 5.9% in the behavior policy. The wiki penalty redistributes a small amount of mass from Question toward Providing Suggestions and Information but does not restore any of the five suppressed strategies, which is consistent with the collapse being intrinsic to the base training rather than to the constraint layer.

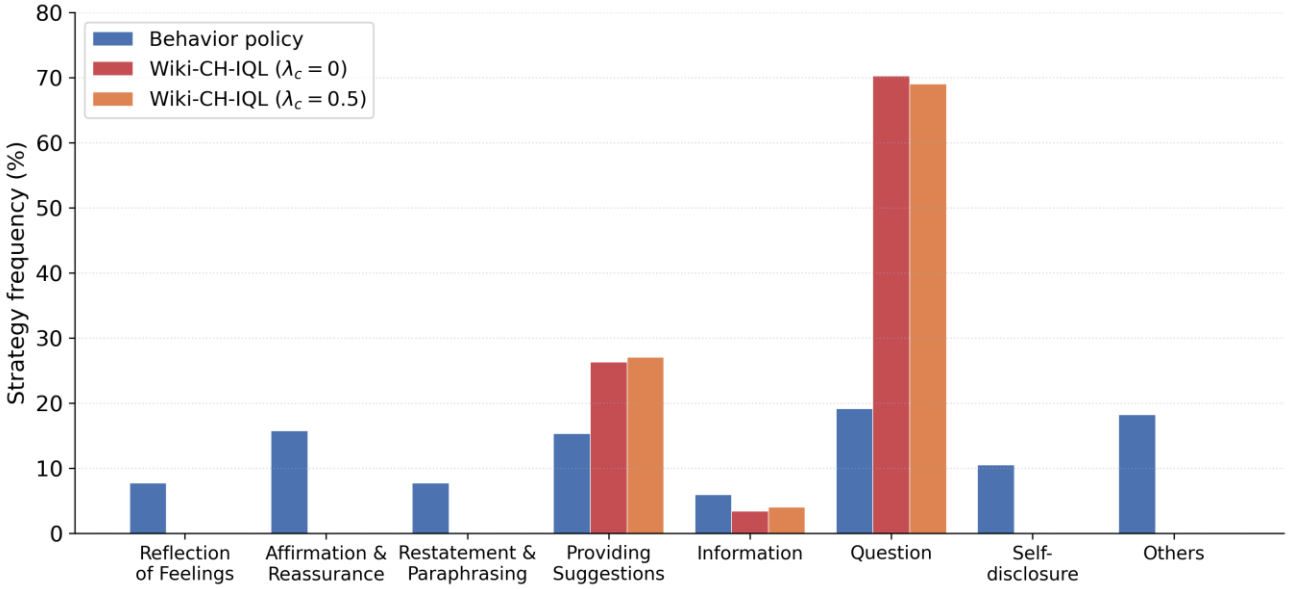

Figure 2. Strategy distribution for behavior policy, unconstrained ablation (λc = 0), and Wiki-CH-IQL (λc = 0.5) on the ESConv test set. Policy collapse to two strategies is present in both policy conditions, confirming it is a base offline RL training issue.

Table 5. Per-strategy frequency on the ESConv test set. Policy collapse to two strategies is identical in both constrained and unconstrained conditions.

Strategy	Behavior (%)	λc=0 (%)	λc=0.5 (%)
Question	19.1	70.2	69.0
Providing Suggestions	15.3	26.3	27.0
Information	5.9	3.4	4.0
Reflection of Feelings	7.7	0.0	0.0
Restatement or Paraphrasing	7.7	0.0	0.0
Self-disclosure	10.5	0.0	0.0
Affirmation and Reassurance	15.7	0.0	0.0
Others	18.2	0.0	0.0

3.3 Dynamic Constraint Update

The DCU experiment modelled a guideline revision in which constraint TC-CI-001 (prohibiting challenge-type strategies under high emotional distress, reflecting trauma-informed care principles) was upgraded from disputed status (conf = 0.50, soft weight) to confirmed status (conf = 0.90, hard weight), without retraining either policy. Figure 3 shows the resulting penalty profiles.

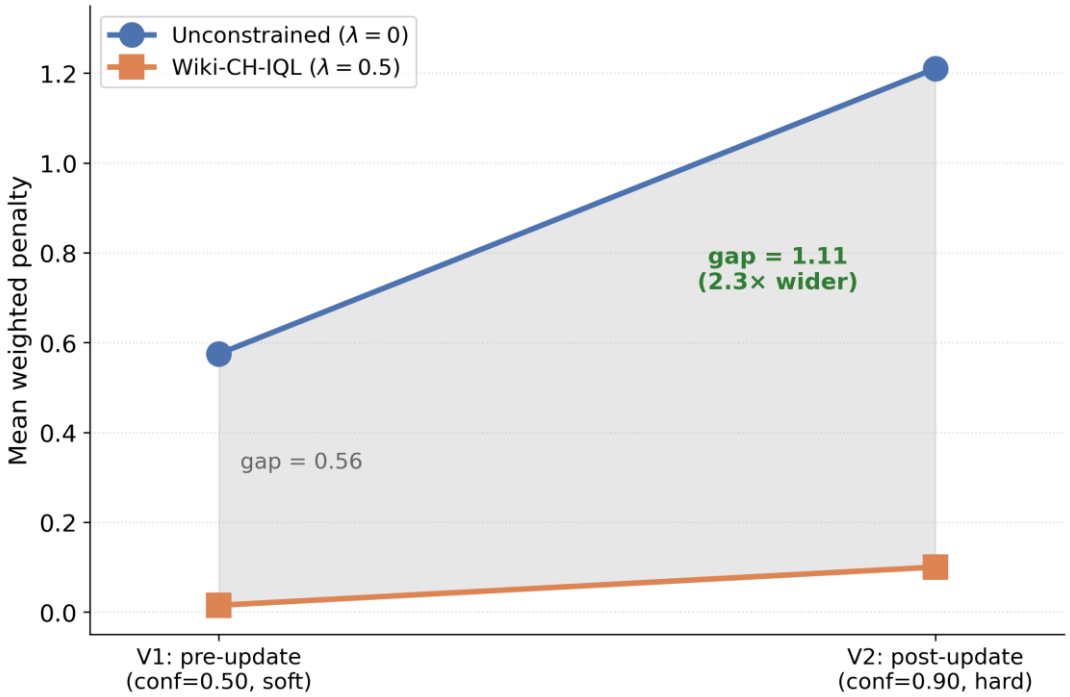

Figure 3. Dynamic Constraint Update experiment. The safety evaluation gap between constrained (λc = 0.5) and unconstrained (λc = 0) policies widens 2.3× after upgrading one constraint from disputed (V1: conf = 0.50, soft) to confirmed (V2: conf = 0.90, hard) status, without model retraining.

The mean weighted penalty gap between the constrained and unconstrained policies widened from 0.56 to 1.11 (ρ = 2.3) without model retraining. This smaller amplification compared to the 7.6× reported in the analogous medical domain experiment [8] is attributable to the lower absolute confidence increment (0.50→0.90 versus 0.50→0.95) and the initial soft-weight assignment producing a smaller per-violation weight increase. The qualitative finding holds: a single wiki revision propagates to a measurable change in the policy safety evaluation gap, without any gradient computation. This is the primary practical contribution of the DCU mechanism for health dialogue systems.

Table 6 summarises the experiment. The revision affects a single wiki page: the confidence of TC-CI-001 rises from 0.50 to 0.90 and its weight class changes from soft (wi = 0.5) to hard (wi = 2.0). Under this single-page edit the mean weighted safety gap between the unconstrained and constrained policies grows from 0.56 to 1.11, an amplification of ρ = 2.3, obtained entirely by re-evaluation of the two fixed policies and without any gradient step, replay of the training data, or change to the network weights.

Table 6. Dynamic Constraint Update result. Upgrading the evidential status of a single constraint page (TC-CI-001) amplifies the policy safety evaluation gap by ρ = 2.3 without model retraining.

Wiki version	TC-CI-001 status	conf	Weight class	Mean weighted safety gap
V1 (pre-update)	Disputed	0.50	Soft (wi = 0.5)	0.56
V2 (post-update)	Confirmed	0.90	Hard (wi = 2.0)	1.11
Amplification	—	—	—	**ρ = 2.3**

4. DISCUSSION

4.1 The DCU Mechanism as the Core Contribution

The most practically significant finding of this study is not the marginal improvement in match rate and macro recall, but the DCU result: revising a single constraint's evidential status amplifies the policy safety evaluation gap by a factor of 2.3 without retraining. From a health informatics systems perspective, this property addresses a fundamental deployment challenge. Clinical guidelines evolve continuously; retraining full RL models each time a guideline is updated is computationally expensive and operationally impractical. The TC-Wiki architecture externalises clinical knowledge from the RL model weights into a maintainable knowledge base, enabling knowledge revision at a fraction of the cost.

This approach parallels retrieval-augmented generation [12] in separating knowledge maintenance from model computation, but provides structured, typed, updatable constraints rather than ad hoc retrieval at inference time. For health dialogue decision support, where clinical safety requirements are both precise and evolving, this architectural property is more valuable than marginal performance gains on a single benchmark.

4.2 Reframing the Narrative: Constraint Control, Not RL Innovation

Following the design principle advocated by Zhang [8], the contribution of this paper is most accurately characterised as a constraint control framework with an RL backbone, not as an RL innovation with a constraint decoration. The CH-IQL algorithm [7] is used essentially unchanged; what is new is the external

knowledge layer that shapes its reward signal in a revision-friendly way. The hyperbolic strategy hierarchy embedding [13], inherited from CH-IQL [7], provides interpretable geometric structure for the strategy space and is retained as the state encoding backbone. It is not claimed as a novel contribution of the present paper.

This reframing has practical implications for evaluation. The small wiki penalty effect (+0.16 pp match rate) is consistent with a soft corrective signal rather than a dominant training force, which is precisely the intended design: the RL dynamics should drive the policy toward expert behaviour while the wiki layer nudges it toward theoretical compliance. Whether a stronger nudge would be beneficial or destabilising requires systematic λc sensitivity experiments deferred to future work.

4.3 Policy Collapse: A Base RL Training Problem

The concentration of 96% of policy outputs on two strategies is a system failure mode that must be clearly distinguished from the wiki constraint contribution. The ablation study confirms it occurs equally without wiki penalties. Contributing factors are: (i) sparse stage-aligned reward providing strong signal for only four of sixteen stage-strategy combinations; (ii) small training set (909 dialogues); and (iii) the simplified hash-based text encoder, which provides less state differentiation than the pre-trained sentence encoder used in CH-IQL [7]. Remedies include conversation-level outcome reward using ESConv post-dialogue survey scores, integration of a pre-trained encoder, and explicit strategy diversity constraints; these directions align with the broader open problems catalogued for offline reinforcement learning [14]. They are the highest-priority directions for future work.

4.4 Limitations

Four limitations merit explicit statement. First, the TC-Wiki constraints were extracted from simulated rather than published clinical texts; the 88.5% executable rate is not a reliable estimate for real counselling manuals and likely overstates the practical constraint yield substantially. Second, the wiki penalty at $\lambda c = 0.5$ contributes marginally to policy metrics, and the optimal λc value is unknown; larger values risk the training instability observed in the medical domain analogue [8]. Third, all evaluation is on held-out transitions from a crowdworker-collected corpus; no assessment with real help-seekers has been conducted. Fourth, the ESConv annotation uses strategy labels assigned by lay supporters, not licensed therapists, which may introduce distributional differences from clinical practice; preference and quality biases in this corpus have also been documented for large language model supporters [15].

This system must not be used as an autonomous therapeutic agent. Any translation toward health care deployment would require: clinician validation of the constraint set by licensed counsellors or psychotherapists; prospective evaluation with real users under ethics board approval; crisis escalation pathways; and privacy-compliant data handling in accordance with applicable regulations.

5. CONCLUSIONS

This proof-of-concept demonstrates that clinical counselling guidance can be partially formalised as a structured, updatable constraint wiki and integrated into offline RL strategy recommendation without architectural change. The controlled ablation confirms that wiki penalties contribute positively but marginally (+0.16 pp match rate, +0.24 pp macro recall) at the current $\lambda c = 0.5$ setting, and that policy collapse is a base RL training problem rather than a constraint artefact. The Dynamic Constraint Update mechanism—enabling single-constraint guideline revisions to propagate to a 2.3× safety gap amplification without retraining—is the primary practical contribution.

The three required next steps are: (i) replacing simulated constraint sources with the actual published clinical manuals; (ii) redesigning the reward signal and state encoder to prevent strategy collapse; and (iii) conducting systematic λc sensitivity analysis to quantify the optimal constraint strength for health dialogue applications.

6. LIST OF ABBREVIATIONS

CH-IQL: Conservative Hyperbolic Implicit Q-Learning; CQL: Conservative Q-Learning; DCU: Dynamic Constraint Update; ESConv: Emotional Support Conversation dataset; IQL: Implicit Q-Learning; LLM: Large Language Model; MI: Motivational Interviewing; RL: Reinforcement Learning; TC-Wiki: Therapeutic Constraint Wiki; W2C: Wiki-to-Constraint.

7. AUTHOR CONTRIBUTIONS

B.Z.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing—original draft, Writing—review and editing. R.Z.: Conceptualization, Validation, Writing—review and editing.

8. FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

9. ETHICS STATEMENT

This study is a computational experiment on a publicly available anonymised corpus (ESConv [11]). No human participants were recruited and no personal data were collected. The ESConv dataset was collected under informed consent procedures described in the original publication [11]. The system described is a research prototype and has not been deployed in any clinical or user-facing context.

10. DATA AVAILABILITY STATEMENT

The ESConv dataset is publicly available at https://github.com/thu-coai/Emotional-Support-Conversation. Code implementing the TC-Wiki W2C pipeline and Wiki-CH-IQL training will be released upon publication. Constraint set and training logs are available from the corresponding author upon reasonable request.

11. ACKNOWLEDGMENTS

ESConv data were provided by the CoAI Group, Tsinghua University. The authors used Claude (Anthropic, claude-sonnet-4-6, 2025) to assist with manuscript preparation and code development; the authors reviewed all content and take full responsibility for the published article.

12. CONFLICT OF INTEREST

The authors declare no conflict of interest.

13. REFERENCES

[1] Hill CE. Helping Skills: Facilitating Exploration, Insight, and Action. 4th ed. Washington, DC: American Psychological Association; 2014.

[2] Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health. 2017;4(2):e19. doi:10.2196/mental.7785

[3] Grodniewicz JP, Hohol M. Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. Front Psychiatry. 2023;14:1190084. doi:10.3389/fpsyt.2023.1190084

[4] Cheng Y, Liu D, Li S, Chen Z, Wang Q, Zhang Y. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In: Proceedings of EMNLP 2022. 2022:3014–3026.

[5] Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit Q-learning. In: International Conference on Learning Representations; 2022.

[6] Kumar A, Zhou A, Tucker G, Levine S. Conservative Q-learning for offline reinforcement learning. Adv Neural Inf Process Syst. 2020;33:1179–1191.

[7] Zhang B, Zhang R. Hierarchy-aware hyperbolic representation with conservative offline policy learning for interpretable strategy recommendation in mental health dialogues. Preprint. 2025.

[8] 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.

[9] Beck AT, Rush AJ, Shaw BF, Emery G. Cognitive Therapy of Depression. New York: Guilford Press; 1979.

[10] Miller WR, Rollnick S. Motivational Interviewing: Helping People Change. 3rd ed. New York: Guilford Press; 2013.

[11] Liu S, Zheng C, Demasi O, Sabour S, Li Y, Yu Z, et al. Towards emotional support dialog systems. In: Proceedings of the 59th Annual Meeting of the ACL; 2021. p. 3469–3483. doi:10.18653/v1/2021.acl-long.269

[12] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020;33:9459–9474.

[13] Nickel M, Kiela D. Poincaré embeddings for learning hierarchical representations. Adv Neural Inf Process Syst. 2017;30.

[14] Levine S, Kumar A, Tucker G, Fu J. Offline reinforcement learning: tutorial, review, and perspectives on open problems. arXiv:2005.01643; 2020.

[15] Kang D, Kim S, Kwon T, Moon S, Cho H, Yu Y, et al. Can large language models be good emotional supporters? Mitigating preference bias on emotional support conversation. In: Proceedings of ACL 2024; 2024. p. 15232–15261. doi:10.18653/v1/2024.acl-long.813