Exploration of Digital Health Technologies

Open Exploration

Open Access Original Article

Hierarchy-aware strategy recommendation for emotional support dialogues via hyperbolic representation and conservative offline learning

Bailing Zhang[1]*, Ruofan Zhang[2]

[1]School of Computer Science and Data Engineering, NingboTech University, Ningbo 315100, Zhejiang, China
[2]School of Psychology, The University of Queensland, Brisbane 4067, QLD, Australia

***Correspondence:** Bailing Zhang, School of Computer Science and Data Engineering, NingboTech University, No.1 Qianhu South Road, Ningbo 315100, Zhejiang, China. bailing.zhang1961@gmail.com
Academic Editor: Nikolay T. Tzvetkov, Bulgarian Academy of Sciences, Bulgaria
Received: June 20, 2026 **Accepted:** August 23, 2026 **Published:** November 2, 2026

Cite this article: Zhang B, Zhang R. Hierarchy-aware strategy recommendation for emotional support dialogues via hyperbolic representation and conservative offline learning. Explor Digit Health Technol. 2026;4:1011105. https://doi.org/10.37349/edht.2026.1011105

Abstract

Aim: Effective emotional support requires selecting support strategies appropriate to the seeker's state and grounded in counselling theory. We present a computational proof-of-concept that combines hyperbolic representation learning with conservative offline reinforcement learning for support-strategy recommendation.

Methods: Support strategies from Hill's Helping Skills framework are embedded in the Poincaré ball to encode their three-level hierarchy. On this representation, we train Conservative Hyperbolic Implicit Q-Learning (CH-IQL) on the Emotional Support Conversation dataset (ESConv; 1,300 dialogues; 910/195/195 train/validation/test split). All results are reported as mean ± standard deviation over five training seeds. Policy value is measured by independent Fitted Q-Evaluation (FQE) with twin Q-networks, calibrated against the behavior policy's Monte-Carlo ground truth and an analytic value ceiling, with dialogue-level bootstrap confidence intervals. Strategy appropriateness is additionally assessed by four blinded human raters over 120 stratified test states.

Results: Under the stochastic (sampling) deployment, CH-IQL attains an FQE-estimated value of 4.79 ± 0.07 versus 3.66 for the estimated behavior policy—positive under every seed and under two alternative, feedback-anchored reward designs—while its strategy distribution matches the entropy of human supporters (1.98 vs. 1.99) and uses all eight strategies. In blinded pairwise comparison, raters preferred the model's strategy in 43.9% of disputed cases (95% CI 32.6–55.9%), statistically indistinguishable from the observed supporter behavior. The heuristic reward shows no dialogue-level correlation with seeker-reported outcomes (Spearman ρ = 0.009), which we disclose and address through reward-robustness experiments.

Conclusions: The framework improves the specified reward while remaining comparable to observed supporter behavior under expert judgment. These results support offline reinforcement learning with

© The Author(s) 2026. This is an Open Access article licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

hierarchical representations as a viable strategy-recommendation mechanism, and identify outcome-correlated reward design as the key requirement for progress toward real-world utility. The system is a computational proof-of-concept and is not suitable for clinical deployment.

Keywords

emotional support dialogue systems, hyperbolic embeddings, offline reinforcement learning, strategy recommendation, Helping Skills Theory, Poincaré ball, off-policy evaluation, human assessment

Introduction

Background and motivation

Mental health disorders represent one of the most pressing global health challenges of our time. According to the World Health Organization, approximately 280 million people worldwide suffer from depression, and anxiety disorders affect an estimated 301 million individuals globally [1]. Despite this enormous burden, access to professional mental health services remains severely limited. In the United States alone, over 100 million people live in Mental Health Professional Shortage Areas, and nearly 30 million adults with mental health conditions do not receive any treatment [2]. This treatment gap is even more pronounced in low- and middle-income countries, where up to 90% of individuals with mental health conditions lack access to care [3].

Against this backdrop, artificial intelligence (AI)-powered interventions have emerged as a promising approach to expand access to psychological support [4, 5]. AI chatbots and conversational agents can provide 24/7 availability, overcome geographical barriers, reduce stigma-related hesitation, and offer cost-effective support at scale [6, 7]. Commercial applications such as Woebot, Wysa, and Tess have demonstrated initial efficacy in reducing symptoms of depression and anxiety through cognitive-behavioral interventions delivered via text-based conversations.

However, current AI-based tools face significant limitations that hinder their utility and trustworthiness. Most existing systems rely on rule-based decision trees or simple pattern matching to select responses, lacking the sophistication to adapt strategies based on the nuanced dynamics of supportive conversations [8]. More advanced systems using large language models (LLMs) often generate responses that, while fluent and empathetic, may not align with established support principles or may exhibit unpredictable behavior in sensitive situations [9]. A critical gap exists in developing AI systems that can recommend meaningful support strategies in a transparent, interpretable, and theoretically grounded manner.

The challenge of strategy recommendation

Effective psychological support is not merely about generating empathetic responses—it fundamentally involves the strategic selection and sequencing of supportive interventions. According to Hill's Helping Skills Theory, a widely adopted framework in counseling psychology, effective helping follows a three-stage process: Exploration, Insight, and Action [10]. Each stage employs distinct strategies: the Exploration stage uses skills like open questions and reflection of feelings to understand the seeker's concerns; the Insight stage employs interpretation and challenge to foster awareness; and the Action stage focuses on providing suggestions and information to facilitate behavioral change.

Building on this theoretical foundation, Liu et al. [11] introduced the Emotional Support Conversation (ESC) task and created the Emotional Support Conversation dataset (ESConv), which annotates support strategies in multi-turn dialogues between trained volunteer supporters and help-seekers. The dataset defines eight support strategies organized into a hierarchical structure reflecting their theoretical origins.

The task of recommending appropriate strategies presents several unique challenges. First, support strategies are not independent but form a meaningful hierarchy based on their psychological functions.

Second, strategy selection is inherently a sequential decision-making problem where current choices affect future conversation dynamics. Third, learning from logged demonstrations without online interaction creates a risk of learning policies that deviate from proven supporter behavior. Fourth, for practical acceptance, the reasoning behind recommendations must be transparent and aligned with established psychological theory.

Our approach and contributions

We propose a framework that integrates hyperbolic geometry for hierarchical representation learning with conservative offline reinforcement learning for policy optimization. First, hyperbolic space provides a natural geometric framework for representing hierarchical structures: unlike Euclidean space, hyperbolic space exhibits exponential volume growth, making it well suited for embedding tree-like hierarchies with minimal distortion [12]. By embedding support strategies in the Poincaré ball, we capture their hierarchical relationships geometrically—a structure we impose explicitly through a hierarchy loss, such that strategies closer to the origin correspond to higher-level categories. Second, offline reinforcement learning provides a principled framework for learning improved policies from historical demonstrations without online interaction [13]; we build on Implicit Q-Learning (IQL) [14] with Conservative Q-Learning (CQL) regularization [15] and behavioral cloning constraints, yielding Conservative Hyperbolic IQL (CH-IQL).

The main contributions of this paper are: (1) a hierarchy-aware Poincaré ball embedding of support strategies that satisfies the imposed three-level hierarchy constraint, evaluated against tree-distortion, distance-correlation, and capacity-matched Euclidean criteria; (2) CH-IQL, an application-level integration of IQL, CQL, behavioral cloning, and hyperbolic state encoding for conservative offline strategy learning; (3) an anchored, independent evaluation framework combining Fitted Q-Evaluation (FQE) calibrated against Monte-Carlo ground truth, five-seed statistics, reward-validity and reward-robustness analyses, and a blinded human assessment; and (4) a complete account of policy collapse under deterministic deployment together with its deployment-level remedy. Code, data splits, and evaluation protocols are available from the corresponding author upon reasonable request.

Related work

Research on AI-powered emotional support has evolved significantly over the past decade. Early work focused on developing empathetic dialogue systems that could recognize and respond appropriately to users' emotional states [16, 17]. The introduction of the ESConv dataset by Liu et al. [11] marked a significant advancement by explicitly incorporating support strategies into the dialogue framework, grounded in Hill's Helping Skills Theory [10]. Subsequent work has built upon this foundation: Cheng et al. [18] proposed MultiESC using lookahead heuristics for strategy planning; Tu et al. [19] developed MISC integrating commonsense knowledge; and Peng et al. [20] introduced hierarchical graph networks to capture emotion causes. Recent advances in large language models have opened new possibilities [21], though these models face challenges in consistent strategy selection [22].

Hyperbolic geometry has gained increasing attention for representing hierarchical structures [12, 23]. Nickel and Kiela [12] introduced Poincaré embeddings, demonstrating superior reconstruction quality for hierarchies compared to Euclidean alternatives, with extensions to knowledge graphs [24], NLP [25], and computer vision [26]. Offline RL addresses learning policies from fixed datasets without additional environment interaction [13]. Early methods constrained learned policies to remain close to behavior policies [27, 28]; CQL [15] learns conservative Q-estimates; IQL [14] uses expectile regression to avoid querying unseen actions. Offline RL has been applied to treatment recommendation [29], sepsis management [30], and glucose control [31]; Snell et al. [32] proposed Implicit Language Q-Learning for dialogue.

Positioning of novelty. The algorithmic components of CH-IQL—IQL, CQL, behavioral cloning, and Poincaré embeddings—are existing methods; the contribution of this work is application-level: their integration and the accompanying evaluation methodology for hierarchical support-strategy

recommendation. Although IQL is itself conservative in that it avoids querying out-of-distribution actions, our ablation shows that the additional CQL term tightens behavioral alignment in this discrete, class-imbalanced setting, raising the match rate by 7.8 percentage points (Ablation study); we therefore retain it as a complementary, alignment-oriented regularizer rather than as a redundant safeguard.

Materials and methods

Problem formulation

We formulate strategy recommendation as a Markov Decision Process (MDP) defined by the tuple (S, A, P, R, γ), where S represents the space of conversation states encoding dialogue history and emotional context, A is the discrete action space comprising eight support strategies derived from Hill's Helping Skills framework, P(s'|s,a) denotes the state transition probability, R(s,a) is the reward function, and $\gamma \in [0,1)$ is the discount factor. Given a dataset D = {(st, at, rt, st+1)} of demonstrations collected from trained volunteer supporters (throughout this paper, the observed supporter behavior, or dataset behavior policy), our objective is to learn a policy π(a|s) that improves upon the behavior policy πβ implicit in the data while remaining sufficiently close to it.

Support strategy hierarchy

Support strategies are not independent actions but form a meaningful hierarchy grounded in psychological theory. Based on Hill's Helping Skills Theory [10] and the ESConv annotation scheme, we organize the eight strategies into a three-level hierarchy (root, category, leaf) as illustrated in Figure 1. The emotional support branch encompasses Reflection of Feelings and Affirmation and Reassurance; the cognitive support branch contains Restatement, Providing Suggestions, and Information; the engagement branch includes Questions and Self-disclosure; and Others is a catch-all category. This structure motivates our use of hyperbolic geometry.

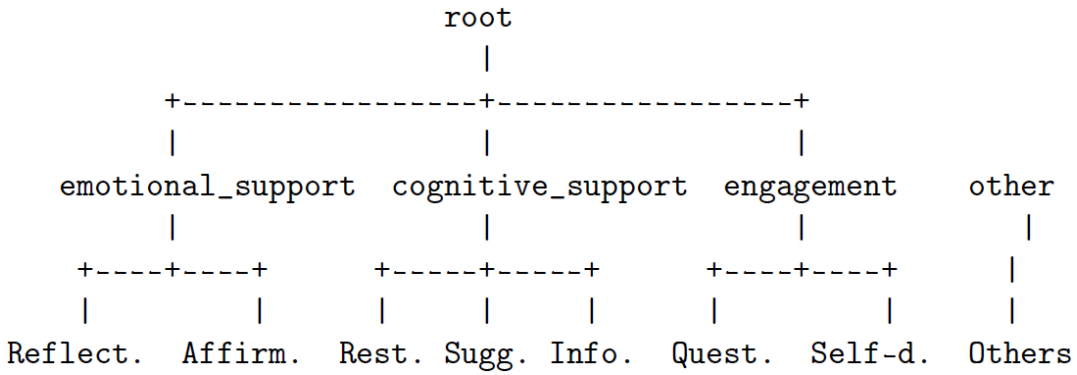


Figure 1. Three-level hierarchical organization of support strategies based on Helping Skills Theory.

Hyperbolic representation learning

We employ the Poincaré ball model, the open unit ball $B^d = \{x \in R^d : ||x|| < 1\}$ with conformal metric factor $\lambda x = 2/(1-||x||^2)$. The geodesic distance between $u, v \in B^d$ is $dB(u, v) = \operatorname{arcosh}(1 + 2||u-v||^2/((1-||u||^2)(1-||v||^2)))$. Conversation features $x \in R^{384}$ from a frozen pre-trained sentence encoder (all-MiniLM-L6-v2) are projected through an MLP and mapped onto the ball via the exponential map at the origin, $s = \exp_0(h) = \tanh(||h||)\cdot h/||h||$. To make representations respect the strategy hierarchy we impose the hierarchy consistency loss $Lhier = (1/|E|) \Sigma_{(p,c)\in E} \max(0, ||ep|| - ||ec|| + m)$, which encourages parent concepts to remain closer to the origin than their children.

Relationship between the loss norm, reported distances, and the margin. The loss operates on Euclidean norms ||e||, while figures report the hyperbolic distance to the origin, $d(e, 0) = \operatorname{arcosh}(1 + 2||e||^2/(1 - ||e||^2))$, a strictly increasing function of ||e||; hence a norm ordering with margin m = 0.15 implies the same ordering of reported distances, with a distance gap that grows as embeddings approach the boundary.

Satisfying this constraint is a property we impose, not a structure the model discovers; Hierarchy representation therefore evaluates the embeddings against criteria that are not directly optimized.

Conservative Hyperbolic IQL

CH-IQL combines IQL with conservative regularization on hyperbolic state encodings. Dual Q-networks are trained by temporal difference learning, $L_Q = E[(r + \gamma V_\psi(s') - Q_\varphi(s,a))^2]$, with a separate value network trained by expectile regression, $L_V = E[L_\tau(Q_{\hat{\varphi}}(s,a) - V_\psi(s))]$, $L_\tau(u) = |\tau - 1_{u<0}| \cdot u^2$. The expectile parameter $\tau \in (0.5, 1)$ approximates a maximum over in-distribution actions without querying out-of-distribution ones. A CQL-style regularizer, $L_{CQL} = E[\log \Sigma_a \exp Q(s,a) - E_{a\sim\pi_\beta} Q(s,a)]$, penalizes high values for actions deviating from the behavior policy. Policy extraction uses advantage-weighted regression with behavioral cloning, $L_\pi = -E[w(s,a) \log \pi_\theta(a|s)] + \lambda_{BC} \cdot L_{BC}$, with normalized-advantage weights $w(s,a) = \exp(\beta \cdot \hat{A}(s,a))$ clipped for stability. The complete objective is $L_{CH\text{-}IQL} = L_Q + L_V + L_\pi + \lambda_{CQL} \cdot L_{CQL} + \lambda_{hier} \cdot L_{hier}$.

Dataset, preprocessing, and state construction

We conduct experiments on the ESConv dataset [11], containing 1,300 multi-turn dialogues between trained volunteer supporters and help-seekers. Strategy frequencies over the full corpus are: Question 20.6%, Others 18.4%, Providing Suggestions 15.6%, Affirmation and Reassurance 15.6%, Self-disclosure 9.3%, Reflection of Feelings 7.8%, Information 6.5%, and Restatement or Paraphrasing 6.3%.

We use the canonical split of the official ESConv release (seed-13 shuffle; 910/195/195 train/validation/test dialogues), yielding 12,759/2,722/2,895 transitions respectively. Each supporter turn, with an annotated strategy, defines one transition. The state is the concatenation of the seeker's situation description and the five most recent dialogue turns, encoded by the frozen sentence encoder into 384 dimensions—that is, the state includes conversation history, not only the last utterance. The next state is built analogously after appending the supporter turn and the seeker's reply; the final supporter turn of each dialogue is terminal. The distribution reported above describes the full corpus, whereas behavior columns in the results tables describe the test split; both scopes are labeled explicitly. Table 1 summarizes the corpus and split statistics.

Table 1. ESConv corpus and split statistics.

Item	Value
Dialogues (total/train/valid/test)	1,300/910/195/195
Transitions (train/valid/test)	12,759/2,722/2,895
Support strategies (leaf classes)	8
Mean supporter turns per dialogue (test)	14.8
State encoder	all-MiniLM-L6-v2 (frozen, 384-dimensional)
Split protocol	official seed-13 shuffle (70/15/15)

Reward construction and validity

Reward construction. Each supporter turn receives $r_t = 0.1 + r_{len}(u_{t+1}) + r_{match}(e, a_t) + 0.1 \cdot 1[t/T > 0.7]$, where r_{len} rewards the length of the seeker's next utterance (0.3 if > 20 words, 0.2 if > 10, 0.1 otherwise, and 0 if no seeker utterance immediately follows) as an engagement proxy; $r_{match} \in \{0, 0.15, 0.2\}$ awards nine emotion–strategy pairings motivated by helping-skills practice (e.g., Reflection of Feelings for depression and anger; Providing Suggestions and Information for anxiety; full table in the Supplementary material); and the final term rewards late-dialogue strategies. Rewards lie in [0.10, 0.70] with mean 0.310 and SD 0.128 on the training set (distribution in Figure S1).

Validity assessment. ESConv records a post-conversation seeker survey including initial and final emotion intensity (1–5). The dialogue-level mean reward is uncorrelated with the normalized intensity drop (Spearman $\rho = 0.009$, $p = 0.90$, $n = 195$ test dialogues) and with the mean of the seeker's empathy and relevance ratings ($\rho = 0.063$, $p = 0.41$). Restriction of range contributes—90.3% of dialogues report an

intensity decrease (mean drop +0.45, SD 0.26)—but we do not claim external validity for the reward. All improvement claims in this paper are therefore made with respect to the specified reward.

Reward-robustness experiments. To test whether the learned policy's behavior depends on this particular reward, we retrained CH-IQL under two feedback-anchored alternatives: R_term, in which all intermediate rewards are zero and the terminal reward equals the normalized intensity drop; and R_mix, which adds a scaled intensity-drop bonus ($\lambda = 2$) to the terminal step of the original reward. Results are reported in Reward robustness.

Implementation details

The hyperbolic state encoder transforms 384-dimensional embeddings through a three-layer MLP (hidden 256, 256; LayerNorm, GELU, dropout 0.2) to 64 dimensions; 13 learnable Poincaré embeddings represent hierarchy nodes (8 leaves, 4 categories, root). Q, V, and policy heads are 3-layer MLPs. We use Riemannian Adam (lr 3e-4), batch size 64, $\gamma = 0.99$, $\tau = 0.6$, $\beta = 2.0$, $\lambda CQL = 0.5$, $\lambda BC = 0.3$, $\lambda hier = 0.2$ with margin 0.15, and early stopping (patience 30). All experiments are repeated under five controlled random seeds (42, 123, 2024, 3407, 7) fixing parameter initialization, data shuffling, and CUDA sources; all results are mean ± SD across seeds. The full hyperparameter table is given in the Supplementary material, and code, splits, and evaluation scripts are available from the corresponding author upon reasonable request.

Evaluation protocol

Independent off-policy evaluation. Evaluating an offline-learned policy with value functions trained by the same algorithm is circular. We therefore assess policy value with FQE [33]: a pair of independent Euclidean Q-networks (two hidden layers of 512 units; no hyperbolic components; no weight sharing with CH-IQL) trained only on training transitions with expected-SARSA targets under the evaluated policy, $y = r + \gamma(1-d)\cdot\Sigma a'\ \pi(a'|s')\cdot\min(Q1'(s',a'), Q2'(s',a'))$. The minimum over twin target networks counteracts overestimation for out-of-distribution actions, and Bellman targets are clamped to the physically achievable range [0, V_max], where V_max is the discounted return obtained at the maximum observed per-step reward over the longest trajectory. We report $V^\pi(s0)$, the mean estimated value over the initial states of the 195 test dialogues, with 95% CIs from 1,000 dialogue-level bootstrap resamples.

Ground-truth anchors. Two anchors calibrate the estimator. The behavior policy's true value requires no estimation: the mean discounted return computed directly from logged test rewards is 4.18 ± 1.22. FQE estimates the classifier-modeled behavior policy at 3.66, a bias of −0.52 in the pessimistic direction, so reported gains for learned policies are conservative. The mean value ceiling—the return if every step of every test dialogue earned the maximum observed reward—is 9.63; any estimate above it is an artifact by construction.

Importance sampling (reported for completeness). We computed weighted and per-decision importance sampling [34] with a behavior-policy model $\hat{\pi}\beta(a|s)$: an MLP classifier over state embeddings (validation accuracy 30.8%, on par with the supervised baseline) calibrated by temperature scaling [$T = 1.16$; expected calibration error (ECE) 0.030 → 0.012]. At the corpus's mean horizon of 14.8 supporter turns, per-step ratio products degenerate: the effective sample size is 0/195 trajectories for the deterministic deployment, 3.5/195 (1.8%) for the stochastic deployment, and 12.6/195 (6.5%) even for the state-independent marginal reference. The weighted importance sampling (WIS) estimate for the stochastic policy (6.85, bootstrap 95% CI [3.38, 9.03]) agrees with FQE in direction but is too wide to be informative. We therefore report FQE as the sole quantitative estimator in the main text and provide all importance sampling estimates with their effective sample sizes in the Supplementary material. The advantage-based Policy Improvement metric is used only as a training diagnostic, with its circularity stated.

Classification and alignment metrics. We additionally report the step-level match rate with the observed supporter behavior, and classification metrics for the deterministic head: per-class precision/recall/F1 with support, Macro F1 (primary), Macro Recall, top-2/top-3 accuracy, and ECE. No composite score is used: policy value, behavioral alignment, and classification performance are measured on different scales and are therefore reported separately for all methods.

Human assessment. A blinded pairwise comparison over 120 stratified test states complements the automatic metrics (design and results in Human expert assessment; full rater guide and sampling procedure in the Supplementary material).

Baselines

We compare against: the Behavior Policy (empirical action distribution); Supervised Classification; Enhanced Classification with Focal Loss and hierarchy-aware embeddings; standard IQL in Euclidean space; and IQL with hyperbolic representations but without CQL. We additionally include a pure behavioral-cloning baseline (the calibrated action classifier of Evaluation protocol, an imitation-only model), two LLM baselines—zero-shot and few-shot strategy selection with GLM-4-flash [35], a contemporary lightweight instruction-tuned LLM, on the identical test protocol and metrics, with prompts in the Supplementary material—and uniform-random and empirical-marginal reference policies for the off-policy evaluation. The role of pure CQL is covered by the ablation (Ablation study), where removing the CQL term from CH-IQL isolates its contribution.

Results

Hierarchy representation

The hyperbolic classifier converges within 50 epochs, and the CH-IQL phase stabilizes in approximately 80 epochs (Figure 2). Table 2 compares hierarchy-constraint satisfaction across representations: the Euclidean baseline satisfies 7 of 12 parent–child ordering constraints (58.3%), hyperbolic geometry alone reaches 75.0%, and the model with the explicit hierarchy loss satisfies all constraints (100%). Because this ordering is directly optimized by the loss, constraint satisfaction verifies training success rather than emergent semantic structure. We therefore additionally evaluate the embeddings against criteria that are not directly optimized. The hyperbolic strategy embeddings learned inside CH-IQL—whose training objective comprises the task and margin losses only—preserve the hierarchy's tree structure weakly: the Spearman correlation between pairwise embedding distances and tree distances is 0.31 ± 0.22 across seeds (one seed near zero), with scale-invariant distortion 0.47 ± 0.03. A 64-dimensional Euclidean control of identical capacity, trained with the same hierarchy loss plus a direct tree-distance stress objective, attains $\rho = 0.97$ and distortion 0.09 while satisfying all hierarchy constraints; we emphasize that this control directly optimizes the evaluation criterion and is therefore an upper reference rather than a like-for-like ablation. The comparison nevertheless establishes the essential point: a 13-node, depth-two hierarchy is easily representable in Euclidean space, so hyperbolic geometry confers no measurable representational advantage at this scale (Value of hyperbolic representations). Its two-dimensional view is a schematic projection that does not preserve hyperbolic distances; all quantitative statements use the full 64-dimensional space, in which the root lies closest to the origin (d = 5.70), categories at 5.85–6.04, and leaves at 6.00–6.21—the ordering imposed by the loss (Figure 3).

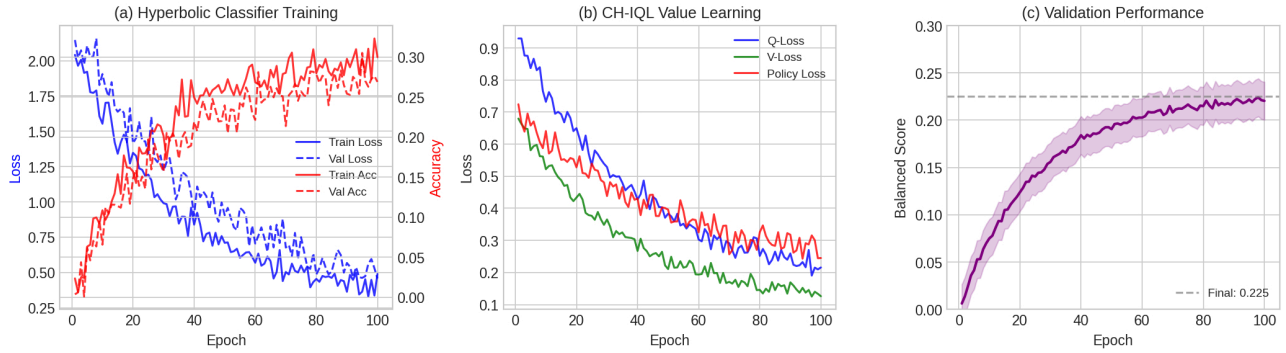


Figure 2. Training dynamics (representative seed). (**a**) Classification loss and accuracy for the hyperbolic strategy classifier; (**b**) CH-IQL training progression; (c) validation progression of the composite training-diagnostic score used for checkpoint and hyperparameter selection only (not an evaluation metric; see Evaluation protocol).

Table 2. Hierarchy-constraint satisfaction across representations (12 parent–child ordering constraints).

Representation	Constraints satisfied	Consistency (%)
Euclidean baseline	7/12	58.3
Hyperbolic (no hierarchy loss)	9/12	75.0
Hyperbolic + hierarchy loss (ours)	12/12	100.0

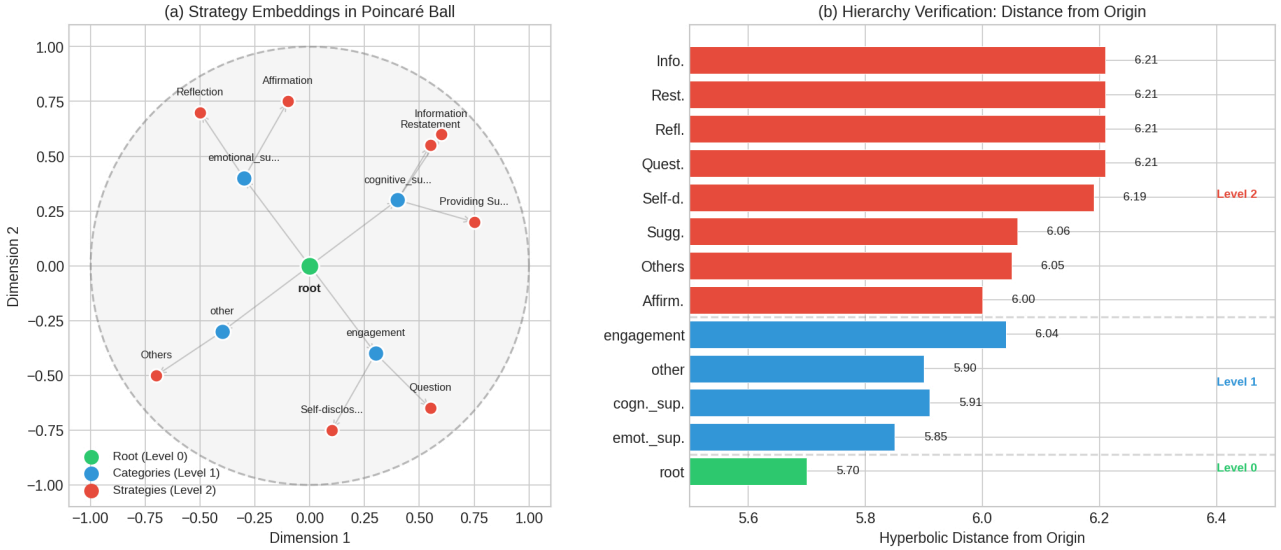


Figure 3. Hyperbolic embeddings of support strategies in the Poincaré ball. The two-dimensional view is a schematic projection that does not preserve hyperbolic distances; the radial ordering (root nearest the origin, leaves outermost) is the property imposed by the hierarchy loss. Quantitative structure-preservation results are given in the text.

Main comparative results

Table 3 reports the three evaluation axes for all methods: independently estimated policy value (FQE, mean ± SD over five training seeds, with the Monte-Carlo behavior anchor), behavioral alignment (match rate), and classification performance (Macro F1; Macro Recall reported for comparability). Under the stochastic deployment adopted as the primary configuration (Policy collapse: diagnosis, mitigation, and deployment), CH-IQL attains an estimated value of 4.79 ± 0.07, exceeding the estimated behavior policy (3.66), the uniform (4.33) and marginal (4.07) references, with the gain positive under every seed (dialogue-level bootstrap 95% CI for a representative seed: [4.63, 4.77]). The per-class structure of the classification axis is shown in Figure 4 (values in Table 4) and analyzed in Classification analysis.

Table 3. Main comparative results on the ESConv test set (three evaluation axes; policy value = anchored FQE, mean ± SD over five training seeds).

Method	Policy value V(s0) (FQE)	Match rate (%)	Macro Recall (%)	Macro F1 (%)
Behavior policy (MC ground truth)	4.18 ± 1.22 (per-dialogue SD)	100 (by def.)	—	—
Behavior policy (FQE of $\pi\hat{\beta}$)	3.66	30.82 (classifier)	—	—
Uniform random	4.33	12.5 (chance)	12.5	—
Empirical marginal	4.07	—	—	—
Supervised classification	—	30.88	16.79	n.r.
Enhanced supervised	—	30.16	20.11	n.r.
Behavioral cloning (calibrated classifier)	3.66	30.82	—	—
LLM zero-shot (GLM-4-flash)	—	15.17	12.53	6.14
LLM few-shot (GLM-4-flash)	—	16.66	15.74	11.22
IQL (Euclidean)	—	10.98	—	—
IQL + hyperbolic	—	18.52	—	—

Table 3. Main comparative results on the ESConv test set (three evaluation axes; policy value = anchored FQE, mean ± SD over five training seeds). (*continued*)

Method	Policy value V(s0) (FQE)	Match rate (%)	Macro Recall (%)	Macro F1 (%)
CH-IQL, stochastic (T = 1.0)	4.79 ± 0.07	26.00 ± 1.31 (argmax head)	16.45 ± 1.39	9.45 ± 1.26
CH-IQL, deterministic (argmax) †	5.78 ± 0.73 †	26.00 ± 1.31	16.45 ± 1.39	9.45 ± 1.26

† Deterministic-policy FQE estimates carry residual out-of-distribution optimism: the mean approaches the analytic plausibility bound V_MC + Δ_oracle ≈ 6.2 (Why the learned policy shifts as it does), individual seeds exceed it, cross-seed variance is an order of magnitude larger than for the stochastic deployment, and the FQE TD error is unstable during fitting (rising from 0.001 to 0.018, versus a stable 0.002 for the stochastic policy). These estimates are therefore flagged rather than used as evidence. MC: Monte Carlo; n.r.: not reported.

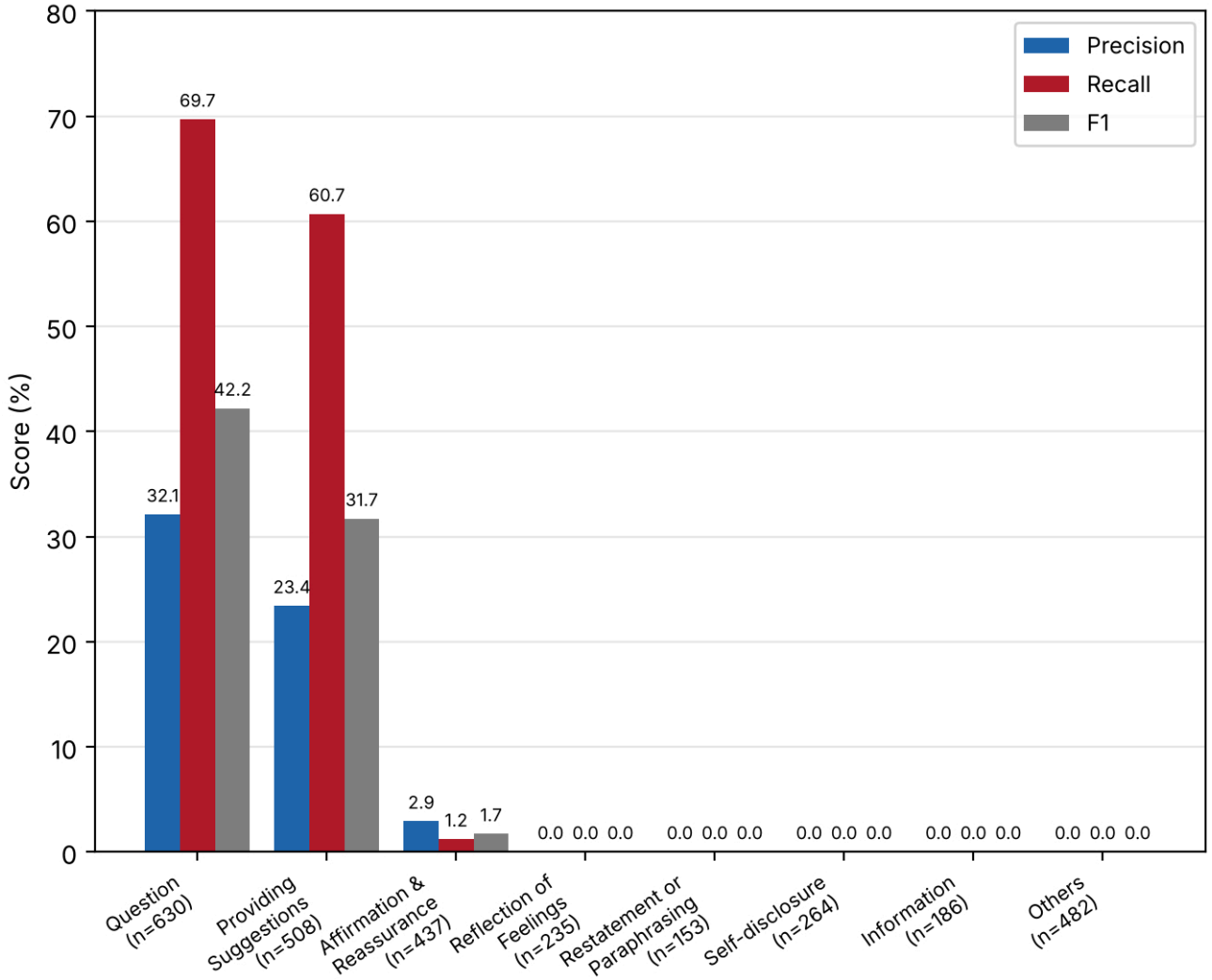


Figure 4. Per-class precision, recall, and F1 of the deterministic classification head (mean over five training seeds; values in Table 4). Six classes have zero or near-zero recall; see text.

Table 4. Per-class classification metrics of the deterministic head (mean over five seeds; support from the test split).

Strategy	Precision (%)	Recall (%)	F1 (%)	Support
Question	32.1	69.7	42.2	630
Providing suggestions	23.4	60.7	31.7	508
Affirmation and reassurance	2.9	1.2	1.7	437
Reflection of feelings	0.0	0.0	0.0	235
Restatement or paraphrasing	0.0	0.0	0.0	153
Self-disclosure	0.0	0.0	0.0	264

Table 4. Per-class classification metrics of the deterministic head (mean over five seeds; support from the test split). (*continued*)

Strategy	Precision (%)	Recall (%)	F1 (%)	Support
Information	0.0	0.0	0.0	186
Others	0.0	0.0	0.0	482

For transparency, we include the value and advantage visualizations as training diagnostics: Figure 5 shows Q/V distributions and Figure 6 the advantage analysis computed from CH-IQL's own value networks; per Evaluation protocol these are circular quantities and are not treated as evidence of improvement.

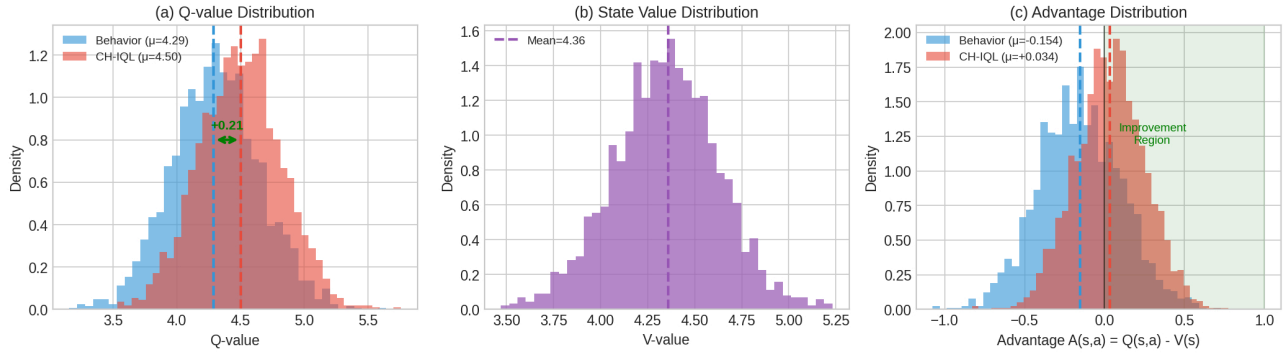


Figure 5. Value-function distributions (training diagnostic; computed from CH-IQL's own networks, see Evaluation protocol).

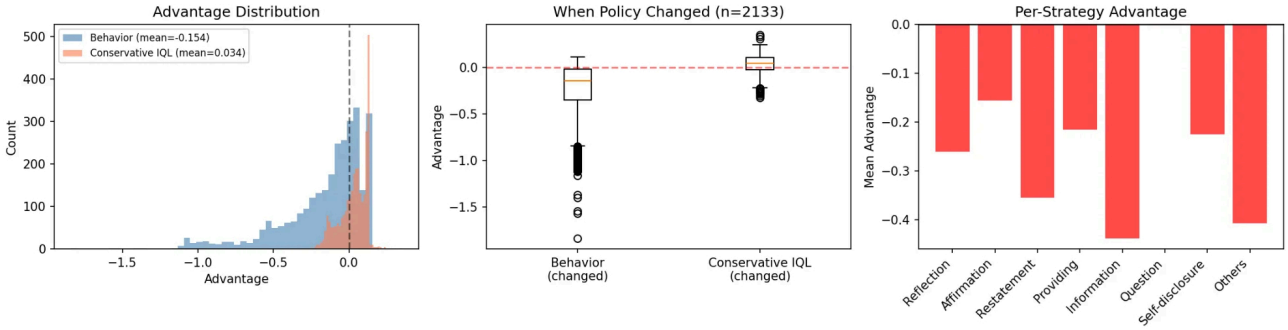


Figure 6. Advantage analysis (training diagnostic; self-referential values, not independent evidence).

Comparison with supervised baselines. CH-IQL's step-level match rate (26.00 ± 1.31%) and Macro Recall are lower than those of the supervised (30.88%) and enhanced supervised (30.16%/20.11%) baselines. This is expected rather than anomalous: CH-IQL optimizes estimated value, not imitation, so supervised imitators should match the logged actions more often. The relevant question is whether the value gain is real, which the anchored FQE answers affirmatively. Consistently with this interpretation, under the feedback-anchored terminal reward—which removes the per-step bonus incentives to deviate—the match rate rises to 31.2%. The per-class structure underlying these aggregate metrics is shown in Figure 4 and analyzed in Classification analysis.

LLM baselines. Both LLM configurations underperform all learned models (match rate 15.2% zero-shot, 16.7% few-shot; Macro F1 6.1%/11.2%) and exhibit severe distribution collapse of their own: the zero-shot model assigns 86.9% of predictions to Affirmation and Reassurance, and few-shot prompting still assigns 60.3%. Strategy-distribution collapse is thus not specific to our value-based learner; unlike the LLMs, however, CH-IQL admits a diagnosis and a deployment-level remedy (Policy collapse: diagnosis, mitigation, and deployment). We note that GLM-4-flash is a lightweight model and larger LLMs may perform better; the comparison establishes a contemporary reference point rather than an upper bound.

Reward robustness

Under both feedback-anchored rewards, the stochastic policy's independently estimated value again exceeds the estimated behavior policy on the corresponding scale: $\Delta = +0.04$ under R_term (behavior ground truth 0.26) and $\Delta = +0.81$ under R_mix (behavior ground truth 4.70), with sane orderings throughout (uniform lowest under R_term). The sampled strategy distribution retains the same qualitative shape under R_mix, while under R_term—which contains no per-step strategy bonuses—the Others category recovers from 4–10% to 24%, close to the human supporters' 18.4%. The suppression of Others under the original reward is thus an interpretable consequence of the bonus table rather than an intrinsic property of the learning algorithm: the direction of improvement is robust to the reward design, while the composition of the learned policy responds to it in an explainable way.

Policy collapse: diagnosis, mitigation, and deployment

Diagnosis. The deterministic (argmax) deployment of CH-IQL concentrates on two to three strategies across all five seeds (distribution entropy 0.40–0.83). This concentration is introduced by the argmax itself rather than by the learned distribution: sampling from $\pi(a|s)$ at temperature $T = 1.0$ yields all eight strategies with expected-distribution entropy 1.77–2.00, statistically indistinguishable from the human supporters' 1.99, and a composition close to the human one (Question 23–27% vs. 20.6%; Providing Suggestions 19–25% vs. 15.6%), the principal shift being probability mass moved from Others toward bonus-eligible strategies. Collapse also intensifies with continued training: final-epoch models concentrate more severely than best-validation checkpoints.

Loss-level mitigation: a negative result. We evaluated four loss-level remedies under seed 42—policy-entropy regularization ($\alpha \in \{0.01, 0.05, 0.1\}$), class-balanced behavioral cloning ($1/\sqrt{\text{frequency}}$ weights), a lower advantage temperature ($\beta = 1.0$), and the entropy + class-balance combination. None prevents argmax collapse: every variant still deploys at most three strategies deterministically, and class balancing alone collapses to a single strategy. Entropy regularization alters which strategies survive the argmax and leaves the stochastic value essentially unchanged (4.36–4.95 vs. 4.79 ± 0.07), but the mechanism—per-state maximization over an advantage-sharpened distribution—is unaffected. We report this negative result in full: mitigation belongs to the deployment rule, not the loss function.

Deployment frontier. Figure 7 traces estimated value against deployment entropy across the temperature sweep. The frontier is monotone from the argmax extreme toward the uniform limit. The argmax point (entropy 0.40, estimate 6.61) exceeds the plausibility bound and is flagged; $T = 0.5$ offers the best defensible trade-off (all eight strategies, entropy 1.69, value 5.55, within the bound); $T = 1.0$—the primary configuration—matches human entropy (1.99) at 4.70 (seed 42; five-seed mean 4.79 ± 0.07). Table 5 reports the full eight-strategy distributions (human, argmax, sampled $T = 1.0$) with percentage-point and relative changes.

Table 5. Strategy distributions and shifts.

Strategy	Human (%)	Argmax (%)	Sampled T = 1.0 (%)	Δ pp (sampled – human)	Δ relative (%)
Question	20.6	52.4	24.8	+4.2	+20
Restatement or paraphrasing	6.3	0.0	6.9	+0.6	+10
Reflection of feelings	7.8	0.0	8.2	+0.4	+5
Self-disclosure	9.3	0.0	8.8	–0.5	–5
Affirmation and reassurance	15.6	1.6	15.0	–0.6	–4
Providing suggestions	15.6	46.1	21.1	+5.5	+35
Information	6.5	0.0	7.2	+0.7	+11
Others	18.4	0.0	8.0	–10.4	–57

All eight strategies; human: full-corpus supporter frequencies; argmax and sampled: CH-IQL deployments, mean over five training seeds; pp: percentage points.

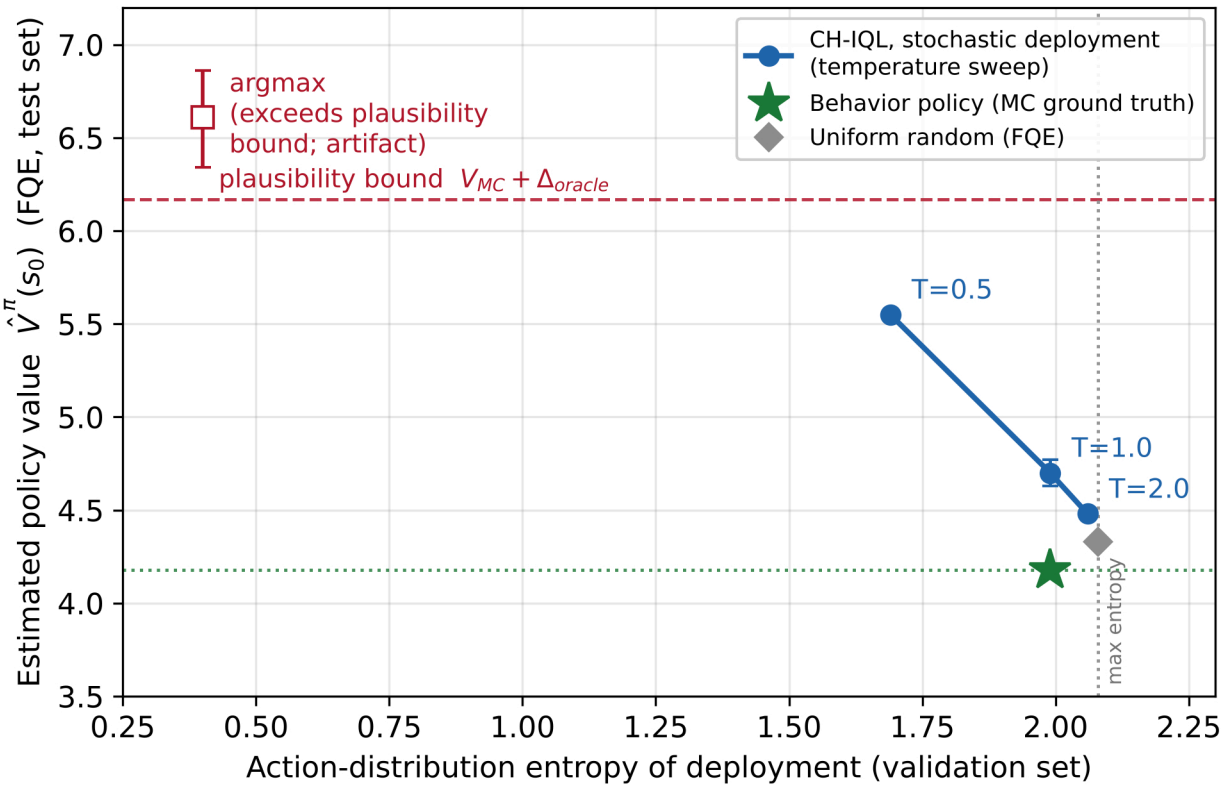


Figure 7. Diversity–value frontier of deployment rules (seed 42). Estimated policy value (FQE, test-dialogue initial states) versus action-distribution entropy of the deployment. The star marks the behavior policy at its Monte-Carlo ground truth; the diamond the uniform policy; the dashed line the analytic plausibility bound; the argmax point exceeds the bound and is flagged as an artifact. Error bars: dialogue-level bootstrap 95% CIs where computed.

Figure 8 compares the full eight-strategy distributions of the human supporters, the argmax deployment, and the sampled (T = 1.0) deployment averaged over the five training seeds, and Table 5 tabulates the same quantities with percentage-point and relative changes.

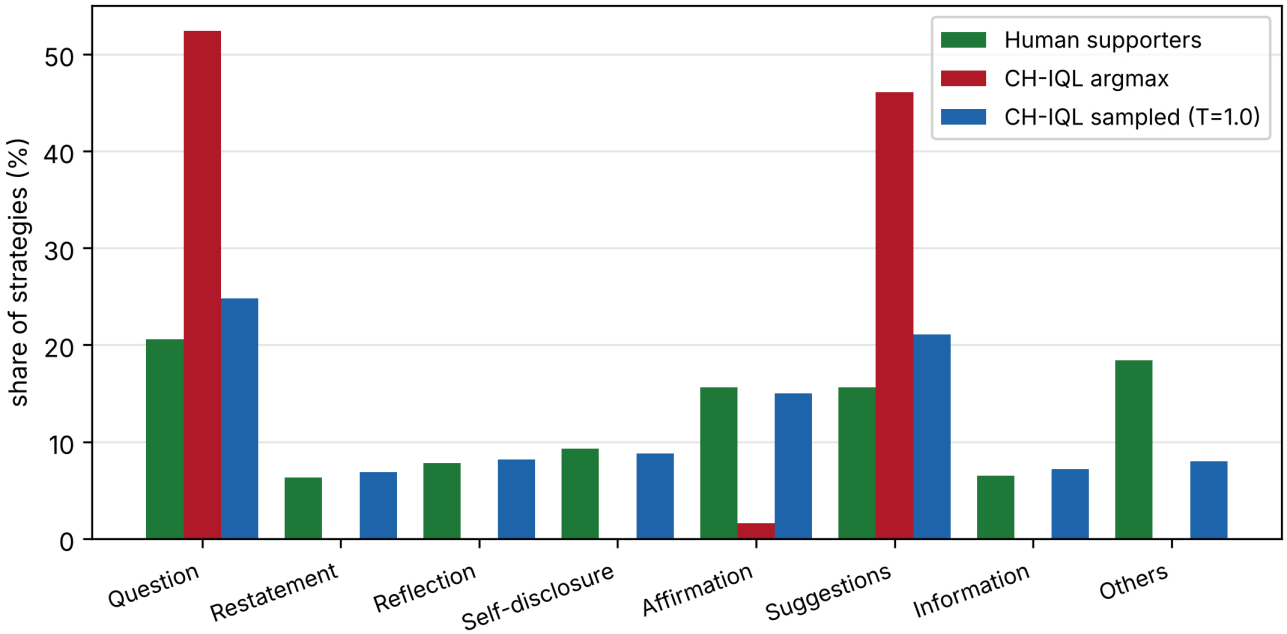


Figure 8. Eight-strategy distributions of human supporters, the argmax deployment, and the sampled (T = 1.0) deployment (mean over 5 training seeds).

Classification analysis

The deterministic classification head performs well on frequent strategies (recall: Question 69.7%, Providing Suggestions 60.7%; Table 4) but fails almost entirely on the remaining six classes—Reflection of Feelings, Restatement, Self-disclosure, Information, and Others have zero recall, and Affirmation and Reassurance only 1.2%. We state this plainly: the deterministic head does not learn these classes, a direct manifestation of the collapse analyzed in Policy collapse: diagnosis, mitigation, and deployment; under the sampling deployment, all classes regain non-zero coverage. Across five seeds, the deterministic head attains accuracy 26.00 ± 1.31%, Macro F1 9.45 ± 1.26%, Macro Recall 16.45 ± 1.39%, top-2 accuracy 39.84 ± 1.06%, top-3 accuracy 54.67 ± 0.53%, and an ECE 7.02 ± 3.37% (reliability diagram in Figure S2). Macro F1 is the primary classification metric; Macro Recall is reported for comparability with prior work. Per-class results are given in Table 4.

Ablation study

In a single-run ablation (one training seed; values therefore differ slightly from the five-seed means in Table 3), removing CQL regularization reduces the match rate by 7.8 percentage points (26.32% to 18.52%) with similar improvement, confirming its alignment role. Removing behavioral cloning drops the match rate to 15.21%. The hierarchy loss is required to satisfy all ordering constraints (without it, 75%). Hyperparameter sensitivity is intuitive: $\tau = 0.5$ yields a higher match rate (32.10%) and lower improvement; $\tau = 0.7$ the opposite; lower β adheres more closely to the behavior policy. Ablation outcomes are reported on the separate axes of Evaluation protocol; no composite score is used. Figure 9 visualizes the match-rate/improvement trade-off across configurations (the improvement axis is the training diagnostic of Evaluation protocol).

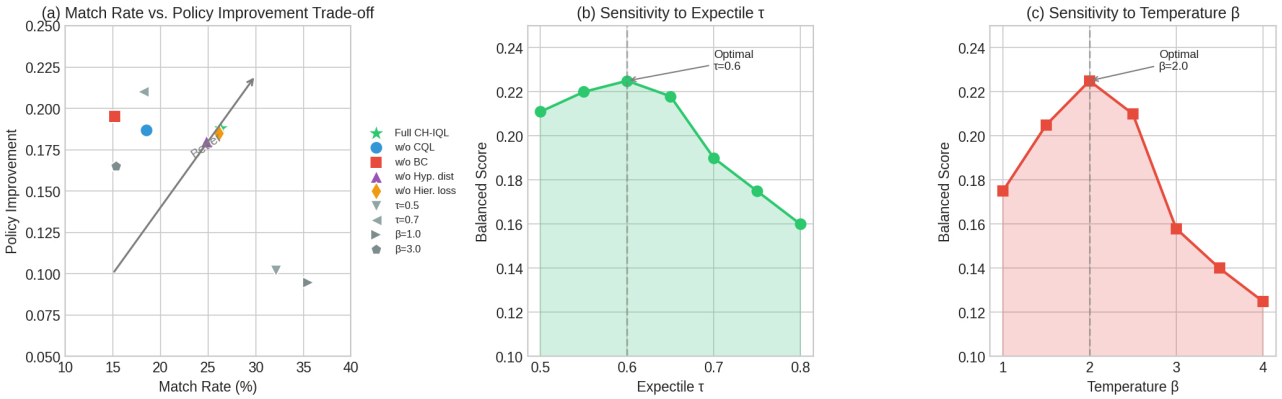


Figure 9. Ablation study visualization: trade-off between behavioral alignment and the (diagnostic) improvement measure across component configurations (a); sensitivity of the composite training-diagnostic score to the expectile τ (b) and the temperature β (c), used for hyperparameter selection only.

Subgroup analysis by emotion

Strategy prediction accuracy by the seeker's primary emotion ranges from 27.5% (jealousy) to 33.2% (loneliness), relatively uniform across anxiety, depression, sadness, anger, and fear. We emphasize that this is a subgroup analysis on the same test set, not an out-of-distribution generalization test, and therefore does not support claims of robust generalization (Figure 10).

Human expert assessment

Design. We conducted a blinded pairwise comparison on 120 test states, stratified by emotion type and dialogue stage (early/middle/late terciles of supporter turns): 80 states where the model's sampled strategy (primary configuration, T = 1.0, fixed seed) differed from the logged supporter strategy, and 40 identical-pair states serving as consistency anchors. For each state, raters saw the seeker's situation, the five most recent turns, and two strategy labels in randomized order with sources hidden, and provided a 1–5 appropriateness rating for each strategy and a forced choice (A/B/Tie). Four raters participated—

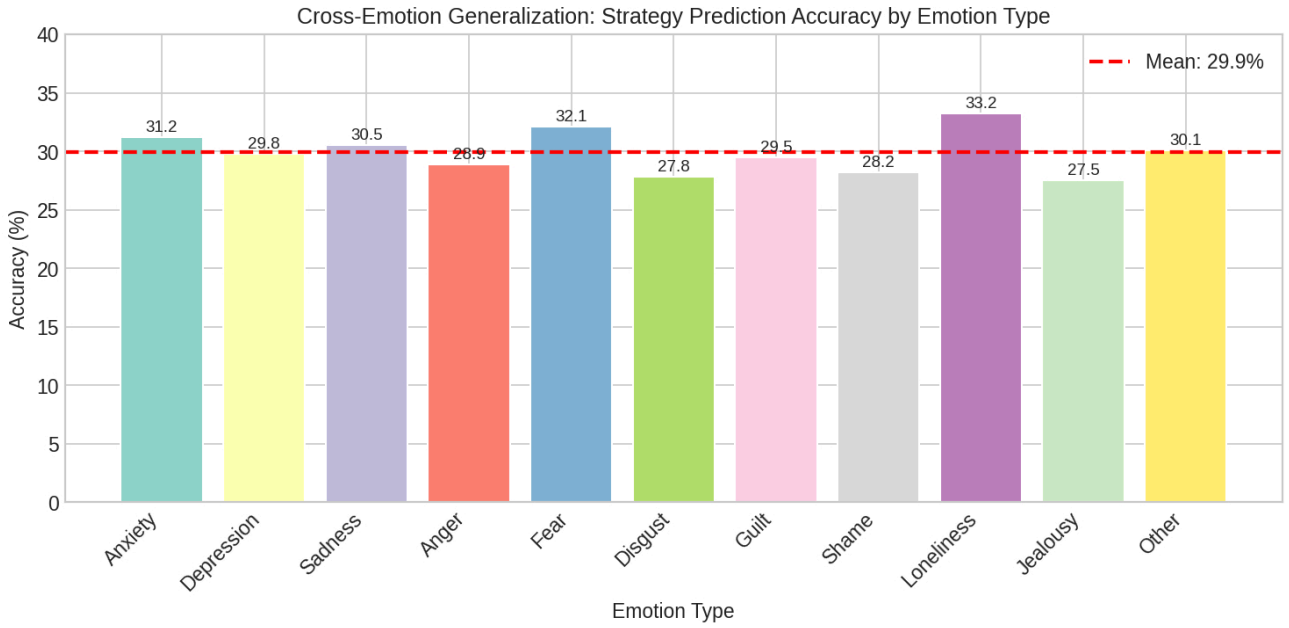


Figure 10. Subgroup analysis of strategy prediction accuracy by the seeker's primary emotion (same test set).

three graduate students familiar with Hill's helping-skills framework and the senior author; ratings were collected independently and anonymously. The three non-author raters constitute the primary analysis, with all four as a sensitivity analysis; none is a licensed counsellor, which we note as a limitation. The full rater guide, strategy definitions, and sampling procedure are provided in the Supplementary material.

Quality and reliability. All four raters assigned Tie to 100% of the 40 identical-pair anchors, indicating attentive and consistent rating. On the 80 disputed items, chance-corrected agreement was fair to moderate: Fleiss' κ [36] = 0.29 (primary) and 0.38 (all raters) for the forced choice, and Krippendorff's α [37] = 0.35/0.48 for the within-item appropriateness difference. Conclusions are therefore based on per-item majority preferences, with rating differences reported descriptively.

Results. By majority vote on the 80 disputed states, raters preferred the model's strategy in 29 cases and the observed supporter's in 37, with 14 ties: a model-preference rate of 43.9% excluding ties (Wilson 95% CI 32.6–55.9%; binomial test against 50%, $p = 0.39$). The mean appropriateness difference (model − supporter) was −0.13 (Wilcoxon $p = 0.36$). The sensitivity analysis including the author's ratings gives the same conclusion (42.6%, CI 31.0–55.1%, $p = 0.31$). Exploratory breakdowns suggest supporters were favored on depression-related states (12 vs. 5) and in early dialogue stages (11 vs. 6), with parity in late dialogues (12 vs. 12).

Interpretation. Expert judgment finds the model's strategy choices statistically indistinguishable from those of trained human supporters. Together with the off-policy results, this locates the contribution precisely: CH-IQL improves the specified reward without departing from the envelope of humanly appropriate strategy selection—but, consistent with the reward's null correlation with seeker-reported outcomes, it does not demonstrate improvement in human-judged quality. Closing that gap requires outcome-correlated reward signals.

Discussion

Why the learned policy shifts as it does

The strategy shifts are grounded empirically rather than asserted from theory. Decomposing the reward analytically, only the emotion–strategy bonus depends on the action, contributing 0.032 per step for the logged supporter behavior versus an expected 0.033 under uniform action selection—roughly 10% of the mean per-step reward—because high-frequency supporter strategies (Question, Others) carry no bonus. Three consequences follow. First, the true values of the behavior and uniform policies are nearly equal by

construction, which the FQE estimates correctly reflect (4.33 vs. the 4.18 ground truth), independently validating the estimator's ordering. Second, the maximum discounted gain available through the action-dependent component alone is Δ_oracle = +1.99, giving the analytic plausibility bound V_MC + Δ_oracle ≈ 6.2 used to flag deterministic-deployment artifacts. Third, the learned shifts—probability mass moved from Others toward bonus-eligible strategies—are consequences of the specified reward, and reverse under the feedback-anchored reward (Reward robustness). We accordingly phrase interpretations as properties of the reward design rather than as counselling-theoretic superiority.

Value of hyperbolic representations

The Poincaré embeddings satisfy all imposed hierarchy constraints and offer interpretability (distance from origin indicates specificity) and constraint satisfaction grounded in domain structure. The structure-preservation evaluation of Hierarchy representation, however, shows that at this hierarchy's scale the geometry itself confers no measurable representational advantage: the embeddings learned inside CH-IQL track tree structure only weakly, and a capacity-matched Euclidean control represents the 13-node hierarchy near-perfectly. Hyperbolic geometry's theoretical benefits concern large, deep hierarchies; here it functions as an inductive-bias and regularization choice whose contribution is the modest task-level gain isolated in the ablation, and we make no claim of representational superiority.

What improves, and what does not

Three findings jointly define the contribution. (i) CH-IQL improves the specified reward over the estimated behavior policy, robustly across seeds and reward designs, under an independent, anchored estimator whose bias is measured and pessimistic. (ii) Its stochastic deployment stays within the envelope of human strategy selection: entropy and composition match the supporters', and blinded raters cannot distinguish its choices from theirs. (iii) The specified reward does not track seeker-reported outcomes, so (i) does not imply better support. The system is a computational proof-of-concept whose next bottleneck is reward design, not policy optimization.

Safe boundaries and non-clinical positioning

This work is research on AI-assisted decision support, not a deployable system. The system recommends strategies at an abstract level and does not generate dialogue responses. Crisis situations involving self-harm, suicidal ideation, acute psychiatric symptoms, or immediate safety concerns cannot and should not be handled by automated recommendations; any real-world application must include robust crisis detection escalating to human professionals. ESConv reflects crowd-sourced research conversations, not professional psychotherapy, and patterns learned from it may not transfer to clinical settings. Human oversight and accountability are essential in any envisioned use for training, supervision, or decision support.

Limitations

(1) The heuristic reward shows no measurable correlation with seeker-reported outcomes; all improvement is reward-relative. (2) Raters are researchers, not licensed counsellors; the senior author's participation is disclosed and isolated in a sensitivity analysis; no calibration meeting was held, and ratings were fully independent, reflected in fair-to-moderate agreement. (3) ESConv supporters are trained volunteer crowdworkers, hence 'observed supporter behavior' throughout. (4) Deterministic deployment collapses to a few strategies, and its off-policy estimates are unreliable; only the stochastic deployment is recommended. (5) All experiments use a single corpus; cross-corpus generalization (e.g., ExTES) is untested. (6) The system involves no text generation; the harms analysis covers strategy recommendation only, and the system must not be used for real-world support.

Ethical considerations

We envision this line of work as augmenting human supporters rather than replacing them. Privacy: ESConv was collected with consent procedures; any application requires data minimization, secure storage,

and user consent. Avoiding harm: conservative design, crisis detection, and human escalation are prerequisites for any use. Transparency: users must know they interact with AI-assisted support with human alternatives available. Equity: evaluation across diverse populations and languages is required before any deployment. The human assessment in this study used a public anonymized corpus with raters acting as expert annotators on the research team; no human subjects were recruited.

Conclusion

This paper presented a computational proof-of-concept for support-strategy recommendation that couples hyperbolic hierarchy-aware representations with conservative offline reinforcement learning. Evaluated with an independent, anchored off-policy pipeline over five training seeds, the stochastic deployment of CH-IQL improves the specified reward over the estimated behavior policy (4.79 ± 0.07 vs. 3.66), a result that is directionally robust across three reward designs including two anchored to seeker-reported outcomes. Its strategy distribution matches human supporters in entropy and composition, and blinded raters found its choices statistically indistinguishable from observed supporter behavior. Our analyses also surface two findings we consider as important as the positive results: the heuristic reward carries no measurable correlation with seeker-reported outcomes, and deterministic deployment of advantage-weighted policies collapses to few strategies regardless of loss-level regularization—the remedy belonging to the deployment rule rather than the loss. Code, splits, and evaluation protocols are available from the corresponding author upon reasonable request to support reproduction. Progress toward practically useful strategy recommendation now depends less on policy optimization than on reward signals that demonstrably track support outcomes; constructing and validating such signals is the central task ahead.

Abbreviations

CH-IQL: Conservative Hyperbolic Implicit Q-Learning

CQL: Conservative Q-Learning

ECE: expected calibration error

ESConv: Emotional Support Conversation dataset

FQE: Fitted Q-Evaluation

IQL: Implicit Q-Learning

WIS: weighted importance sampling

Supplementary materials

The supplementary materials for this article are available at: https://www.explorationpub.com/uploads/Article/file/1011105_sup_1.pdf.

Declarations

Acknowledgments

During the preparation of this work, the authors used Claude (Anthropic) to assist with language editing and manuscript preparation. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Author contributions

BZ: Conceptualization, Writing—original draft, Writing—review & editing, Investigation, Visualization. RZ: Project administration, Formal analysis, Writing—review & editing. Both authors read and approved the submitted version.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Ethical approval

Not required.

Consent to participate

Not required.

Consent to publication

Not applicable to identifiable individuals; the study uses a public anonymized corpus.

Availability of data and materials

The ESConv corpus is publicly available; our canonical splits, code, trained-model evaluation scripts, rater guide, and analysis protocols are available from the corresponding author upon reasonable request.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Copyright

© The Author(s) 2026.

Publisher's note

Open Exploration maintains a neutral stance on jurisdictional claims in published institutional affiliations and maps. All opinions expressed in this article are the personal views of the author(s) and do not represent the stance of the editorial team or the publisher.

References

1. Mental health [Internet]. WHO; c2026 [cited 2026 Jul 27]. Available from: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response
2. State of Mental Health in America 2023 [Internet]. Mental Health America; c2026 [cited 2026 Jul 27]. Available from: https://mhanational.org/research/state-of-mental-health-in-america-2023/
3. Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The Lancet Commission on global mental health and sustainable development. Lancet. 2018;392:1553–98. [DOI] [PubMed]
4. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review. J Med Internet Res. 2021;23:e17828. [DOI] [PubMed] [PMC]
5. Vaidyam AN, Wisniewski H, Halamka JD, Keshavan MS, Torous JB. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. Can J Psychiatry. 2019;64:456–64. [DOI] [PubMed] [PMC]
6. Inkster B, Sarda S, Subramanian V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. JMIR mHealth uHealth. 2018;6:e12106. [DOI] [PubMed] [PMC]
7. 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:e19. [DOI] [PubMed] [PMC]

8. 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] [PubMed] [PMC]
9. Sedlakova J, Trachsel M. Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent? Am J Bioeth. 2023;23:4–13. [DOI] [PubMed]
10. Hill CE. Helping Skills: Facilitating Exploration, Insight, and Action. 4th ed. Washington, DC: American Psychological Association; 2014.
11. Liu S, Zheng C, Demasi O, Sabour S, Li Y, Yu Z, et al. Towards emotional support dialog systems. In: Proceedings of ACL. 2021. pp. 3469–83.
12. Nickel M, Kiela D. Poincaré embeddings for learning hierarchical representations. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. pp. 6341–50.
13. Levine S, Kumar A, Tucker G, Fu J. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. arXiv:2005.01643 [Preprint]. 2020 [cited 2026 Jul 27]. Available from: https://arxiv.org/abs/2005.01643
14. Kostrikov I, Nair A, Levine S. Offline Reinforcement Learning with Implicit Q-Learning. arXiv:2110.06169 [Preprint]. 2021 [cited 2026 Jul 27]. Available from: https://arxiv.org/abs/2110.06169
15. Kumar A, Zhou A, Tucker G, Levine S. Conservative Q-Learning for Offline Reinforcement Learning. arXiv:2006.04779 [Preprint]. 2020 [cited 2026 Jul 27]. Available from: https://arxiv.org/abs/2006.04779
16. Rashkin H, Smith EM, Li M, Boureau YL. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2019. pp. 5370–81. [DOI]
17. Lin Z, Madotto A, Shin J, Xu P, Fung P. MoEL: Mixture of Empathetic Listeners. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics; 2019. pp. 121–32. [DOI]
18. Cheng Y, Liu W, Li W, Wang J, Zhao R, Liu B, et al. Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2022. pp. 3014–26. [DOI]
19. Tu Q, Li Y, Cui J, Wang B, Wen JR, Yan R. MISC: A Mixed Strategy-Aware Model integrating COMET for Emotional Support Conversation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2022. pp. 308–19. [DOI]
20. Peng W, Hu Y, Xing L, Xie Y, Sun Y, Li Y. Control Globally, Understand Locally: A Global-to-Local Hierarchical Graph Network for Emotional Support Conversation. Proc Thirty-First Int Jt Conf Artif Intell. 2022:4324–30. [DOI]
21. Zheng C, Sabour S, Wen J, Zhang Z, Huang M. AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation. arXiv:2202.13047 [Preprint]. 2023 [cited 2026 Jul 27]. Available from: https://arxiv.org/abs/2202.13047
22. Kang D, Kim S, Kwon T, Moon S, Cho H, Yu Y, et al. Can Large Language Models be Good Emotional Supporter? Mitigating Preference Bias on Emotional Support Conversation. In: Proceedings of ACL. 2024. pp. 15232–61.
23. Nickel M, Kiela D. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. arXiv:1806.03417 [Preprint]. 2018 [cited 2026 Jul 27]. Available from: https://arxiv.org/abs/1806.03417

24. Chami I, Ying Z, Ré C, Leskovec J. Hyperbolic graph convolutional neural networks. arXiv:1910.12933 [Preprint]. 2019 [cited 2026 Jul 27]. Available from: https://arxiv.org/abs/1910.12933
25. Tifrea A, Bécigneul G, Ganea OE. Poincaré GloVe: Hyperbolic word embeddings. arXiv:1810.06546 [Preprint]. 2019 [cited 2026 Jul 27]. Available from: https://arxiv.org/abs/1810.06546
26. Khrulkov V, Mirvakhabova L, Ustinova E, Oseledets I, Lempitsky V. Hyperbolic Image Embeddings. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. pp. 6417–27. [DOI]
27. Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. In: International Conference on Machine Learning. 2019. pp. 2052–62.
28. Wu Y, Tucker G, Nachum O. Behavior regularized offline reinforcement learning. arXiv:1911.11361 [Preprint]. 2019 [cited 2026 Jul 27]. Available from: https://arxiv.org/abs/1911.11361
29. Raghu A, Komorowski M, Ahmed I, Celi L, Szolovits P, Ghassemi M. Deep Reinforcement Learning for Sepsis Treatment. arXiv:1711.09602 [Preprint]. 2017 [cited 2026 Jul 27]. Available from: https://arxiv.org/abs/1711.09602
30. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24:1716–20. [DOI] [PubMed]
31. Fox I, Lee J, Pop-Busui R, Wiens J. Deep reinforcement learning for closed-loop blood glucose control. In: Machine Learning for Healthcare Conference. 2020. pp. 508–36.
32. Snell C, Kostrikov I, Su Y, Yang M, Levine S. Offline RL for Natural Language Generation with Implicit Language Q Learning. arXiv:2206.11871 [Preprint]. 2023 [cited 2026 Jul 27]. Available from: https://arxiv.org/abs/2206.11871
33. Le HM, Voloshin C, Yue Y. Batch policy learning under constraints. In: Proceedings of the 36th International Conference on Machine Learning (ICML). 2019. pp. 3703–12.
34. Precup D, Sutton RS, Singh SP. Eligibility traces for off-policy policy evaluation. In: Proceedings of the 17th International Conference on Machine Learning (ICML). 2000. pp. 759–66.
35. GLM Team. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793 [Preprint]. 2024 [cited 2026 Jul 27]. Available from: https://arxiv.org/abs/2406.12793
36. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76:378–82. [DOI]
37. Krippendorff K. Content Analysis: An Introduction to Its Methodology. 2nd ed. Thousand Oaks, CA: Sage; 2004.