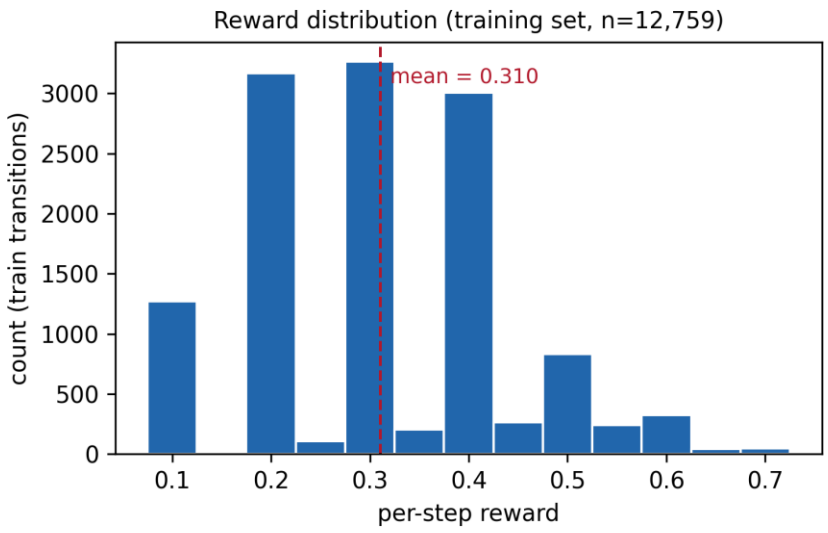


Figure S1. Distribution of the heuristic per-step reward on the training set.

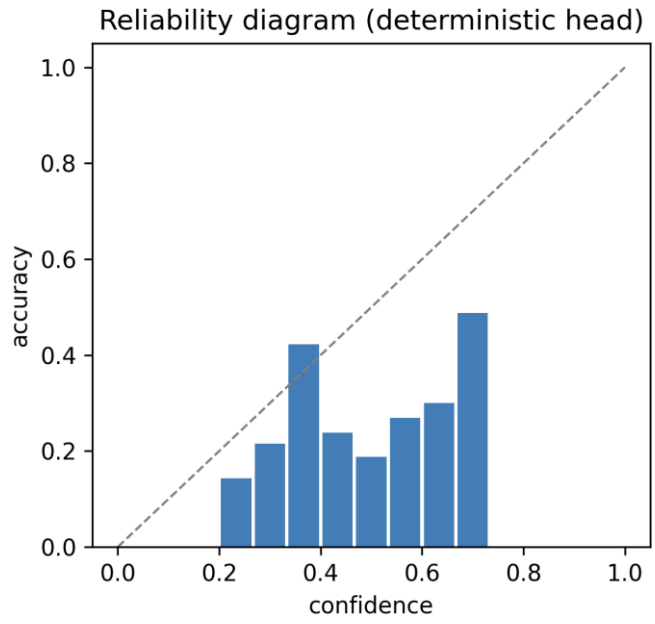


Figure S2. Reliability diagram of the deterministic classification head (representative seed).

Supplementary materials: detailed methods, data, and protocols

Section S1. Reward function and emotion–strategy bonus table

Each supporter turn t with an annotated strategy a_t receives the reward r_t = 0.1 + r_len(u_{t+1}) + r_match(e, a_t) + 0.1·1[t/T > 0.7], where T is the number of supporter turns in the dialogue. The four components are: (i) a constant base reward of 0.1; (ii) an engagement proxy r_len based on the length of the seeker's next utterance u_{t+1} (0.3 if > 20 words, 0.2 if > 10, 0.1 otherwise, and 0 if no seeker utterance immediately follows); (iii) an emotion–strategy bonus r_match(e, a_t), which depends on the seeker's annotated primary emotion e and the chosen strategy and is listed in Table S1 (all other emotion–strategy pairs receive 0); and (iv) a late-dialogue bonus of 0.1 for turns in the final 30% of the dialogue. Only component (iii) depends on the action. Rewards lie in [0.10, 0.70] with mean 0.310 and SD 0.128 on the training set (Figure S1). The final supporter turn of each dialogue is terminal.

Table S1. Emotion–strategy bonus r_match(e, a). Pairs not listed receive 0.

Seeker emotion (e)	Strategy (a)	Bonus
Depression	Reflection of Feelings	0.20
Depression	Affirmation and Reassurance	0.20
Anxiety	Providing Suggestions	0.15
Anxiety	Information	0.15
Anger	Reflection of Feelings	0.20
Sadness	Affirmation and Reassurance	0.20
Fear	Information	0.15
Shame	Affirmation and Reassurance	0.20
Guilt	Reflection of Feelings	0.15

Feedback-anchored reward variants used in the reward-robustness experiments: R_term sets all intermediate rewards to zero and assigns, at the terminal step, the seeker's normalized emotion-intensity drop from the post-conversation survey; R_mix keeps the reward above and adds $\lambda = 2$ times the normalized intensity drop at the terminal step.

Section S2. Hyperparameters

Table S2. Hyperparameters and implementation settings.

Component	Setting
State representation	Seeker situation + five most recent dialogue turns, encoded by frozen all-MiniLM-L6-v2 (384-d)
Hyperbolic state encoder	MLP 384→256→256→64 (LayerNorm, GELU, dropout 0.2); exponential map at the origin onto the Poincaré ball
Hierarchy embeddings	13 learnable Poincaré embeddings (8 leaves, 4 categories, root)
Q, V, and policy heads	3-layer MLPs; dual Q-networks
Optimizer	Riemannian Adam, learning rate 3e-4, batch size 64
Discount factor γ	0.99
Expectile τ	0.6
Advantage temperature β	2.0 (advantage weights clipped)
Loss weights	λCQL = 0.5, λBC = 0.3, λhier = 0.2; hierarchy margin m = 0.15
Early stopping	Patience 30 epochs (best-validation checkpoint retained)
Training seeds	42, 123, 2024, 3407, 7 (initialization, data shuffling, CUDA)
Deployment	Stochastic sampling at T = 1.0 (primary); argmax reported for comparison
FQE	Twin Euclidean Q-networks (two hidden layers of 512 units; no weight sharing with CH-IQL); expected-SARSA targets with minimum over twin targets; Bellman targets clamped to [0, V_max]; trained on training transitions only
Confidence intervals	1,000 dialogue-level bootstrap resamples of the 195 test dialogues
Behavior-policy model	MLP classifier over state embeddings; temperature scaling (T = 1.16)
Loss-level mitigation variants (seed 42)	Policy-entropy regularization $\alpha \in \{0.01, 0.05, 0.1\}$; class-balanced BC ($1/\sqrt{\text{frequency}}$ weights); β = 1.0; entropy (α = 0.05) + class balance
Euclidean control (Section S7)	13 × 64 Euclidean embeddings; Adam, learning rate 1e-2, 2,000 steps; loss = tree-distance stress + 0.2 × hierarchy loss (m = 0.15)

Section S3. Behavior-policy calibration and importance-sampling estimates

The behavior-policy model $\hat{\pi}\beta(a|s)$ used for importance sampling and for the FQE behavior reference is an MLP classifier over state embeddings. Its validation accuracy is 30.8%, on par with the supervised baseline. Temperature scaling (T = 1.16) reduces the expected calibration error from 0.030 to 0.012.

Table S3 reports the importance-sampling estimates on the 195 test trajectories (training seed 42). Because the mean horizon is 14.8 supporter turns, products of per-step importance ratios degenerate, and the effective sample size (ESS) is below 7% of trajectories for every evaluated policy, including the state-independent marginal reference. For this reason FQE is used as the sole quantitative estimator in the main text.

Table S3. Importance-sampling estimates and effective sample sizes (test set, 195 trajectories, training seed 42).

Policy	WIS (95% bootstrap CI)	PDIS	ESS (% of trajectories)
CH-IQL, deterministic (argmax)	Undefined (all trajectory weights zero)	Numerically degenerate	0.0 (0%)
CH-IQL, stochastic (T = 1.0)	6.85 [3.38, 9.03]	—	3.5 (1.8%)
Empirical marginal	—	—	12.6 (6.5%)
Uniform random	3.40 [3.11, 3.71]	1.91	2.9 (1.5%)

WIS: weighted importance sampling; PDIS: per-decision importance sampling; ESS: effective sample size. —: not informative at the observed ESS and not reported.

For reference, Table S4 lists the deployment points of the diversity–value frontier (Figure 7 of the main text) and Table S5 the policy values under the two feedback-anchored rewards.

Table S4. Diversity–value frontier of deployment rules (training seed 42).

Deployment	Action-distribution entropy (validation)	FQE value V(s0) (test)	Note
CH-IQL argmax	0.40	6.61 [6.34, 6.86]	Exceeds plausibility bound (≈ 6.2); flagged
CH-IQL sampled, T = 0.5	1.69	5.55	All 8 strategies used
CH-IQL sampled, T = 1.0	1.99	4.70 [4.63, 4.77]	Primary configuration
CH-IQL sampled, T = 2.0	2.06	4.48	
Behavior policy (MC ground truth)	1.99	4.18	Anchor
Uniform random	2.08	4.33	Reference

Brackets: dialogue-level bootstrap 95% CI where computed. MC: Monte Carlo.

Table S5. FQE policy values under the feedback-anchored rewards (training seed 42).

Policy	R_term	R_mix
Behavior policy, MC ground truth	0.259	4.695
Behavior policy, FQE of $\hat{\pi}\beta$	0.195	4.624
CH-IQL sampled (T = 1.0)	0.233	5.429
CH-IQL argmax	0.180	6.949
Empirical marginal	0.149	4.809
Uniform random	0.124	5.156

Values are on the scale of each reward and are not comparable across columns.

Section S4. Human assessment: rater guide, strategy definitions, and sampling procedure

Sampling procedure. Candidate items were drawn from the 195 test dialogues. Each candidate state was assigned to a stratum defined by the seeker's primary emotion and the dialogue stage (early, middle, or late tercile of supporter-turn position). For each state, the model's strategy was obtained by sampling from the primary configuration (T = 1.0) with a fixed random seed. Eighty disputed items (model strategy different from the logged supporter strategy) and 40 identical-pair anchor items (model strategy equal to the logged strategy) were then drawn with allocation proportional to stratum size, using a fixed random seed. For each item, the left/right (A/B) order of the two strategy labels was randomized independently, and the answer key was stored separately from the rater workbooks.

Rater guide (as provided to raters). Purpose: we are evaluating whether an AI system selects appropriate support strategies in emotional support conversations. You will not see any AI-generated text. Each item shows a real conversation excerpt from ESConv (a public, anonymized corpus of emotional support conversations between trained volunteer supporters and help-seekers) together with two candidate strategy labels for the supporter's next turn. One label is what the human supporter actually did; the other is the system's recommendation. You will not be told which is which, and their order is randomized per item.

Task per item. (1) Read the seeker's background situation and the most recent turns. (2) Read Strategy A and Strategy B. (3) Rate the appropriateness of each strategy independently on the 1–5 scale below: "If the supporter's next utterance used this strategy, how appropriate would that be at this point in the conversation?" (4) Choose which strategy is more appropriate (A/B/Tie); use Tie only when you genuinely cannot rank them, and if both are poor, pick the less inappropriate one unless they are truly equal. (5) Optionally add a short comment, especially when you rate either strategy 1–2 or the choice was difficult. Judge fit to this moment in this conversation, not which strategy is generally nicer.

Table S6. Strategy definitions given to raters (Hill's Helping Skills, as annotated in ESConv).

Strategy	Definition	Example
Question	Asking for information to explore the seeker's situation or feelings	"Can you tell me more about what happened at work?"
Restatement or Paraphrasing	Rephrasing the seeker's own words to show understanding of the content	"So your manager changed your shifts without discussing it with you first."
Reflection of Feelings	Naming or mirroring the seeker's emotional state	"It sounds like you're feeling really overwhelmed by all of this."
Self-disclosure	The supporter shares a relevant personal experience or feeling	"I went through something similar when I lost my job last year."
Affirmation and Reassurance	Affirming the seeker's strengths or providing reassurance and encouragement	"You've handled an incredibly hard situation with real strength."
Providing Suggestions	Suggesting concrete actions or ways of coping	"Maybe you could try writing your worries down before bed."
Information	Providing factual information, resources, or psycho-education	"Many universities offer free counselling services to students."
Others	Utterances outside the above: greetings, small talk, closing pleasantries	"Hello! How has your day been so far?"

Table S7. Appropriateness scale.

Score	Anchor
5	Clearly appropriate: a skilled supporter would very plausibly do exactly this now
4	Appropriate: a reasonable, fitting choice
3	Acceptable: neither helps nor harms the conversation's direction

2	Somewhat inappropriate: mistimed or misfitted, though not harmful
1	Clearly inappropriate: would likely disrupt rapport or the seeker's process

Judging hints. Early in a conversation, exploration and rapport (Question, Restatement, Reflection of Feelings) usually fit, and advice offered before the problem is understood is typically mistimed. Suggestions and Information land better once the seeker's situation and feelings have been acknowledged. Repetition matters: a Question can be ideal, but a fifth consecutive question may feel like an interrogation. Self-disclosure is powerful in moderation. Late-stage turns often support encouragement, summarizing, practical planning, or a warm close. These are tendencies, not rules; the specific context always takes precedence.

Independence and wellbeing. Raters worked independently, rated items in order, did not discuss items with one another, and were asked not to guess which label was the system's. Some conversations discuss distressing topics (e.g., depression, job loss, bereavement); raters could skip any item or pause the task at any time. No pilot calibration meeting was held; all ratings were collected fully independently (see Limitations in the main text).

Section S5. LLM baseline prompts and prediction distributions

Both LLM baselines used GLM-4-flash on the identical test transitions (n = 2,894 parsed items; no unparsable responses). The system prompt was:

"You are an expert in emotional support conversations, familiar with Hill's Helping Skills framework. Given the seeker's situation and the recent dialogue, choose the SINGLE most appropriate support strategy for the supporter's NEXT turn, from exactly these eight options: 0. Question; 1. Restatement or Paraphrasing; 2. Reflection of feelings; 3. Self-disclosure; 4. Affirmation and Reassurance; 5. Providing Suggestions; 6. Information; 7. Others. Answer with the option number only (0–7). No explanation."

The user message contained the state text in the format "[Situation] … [User] … [Supporter] …". In the few-shot condition, three demonstration pairs preceded the test item: (i) "[Situation] I just lost my job and I'm scared about paying rent. [User] I don't even know where to start." → 0 (Question); (ii) "[Situation] My mother passed away last month.

[User] Some days I can't stop crying. [Supporter] That is completely understandable. [User] I just feel so alone in this." → 2 (Reflection of feelings); (iii) "[Situation] I keep procrastinating on my thesis. [User] I know what I should do, I just cannot make myself start." → 5 (Providing Suggestions). The first digit 0–7 in the response was taken as the prediction.

Table S8. LLM baseline results and prediction distributions (%).

	Zero-shot	**Few-shot**
Match rate	15.17	16.66
Macro Recall	12.53	15.74
Macro F1	6.14	11.22
Pred.: Question	3.5	0.1
Pred.: Restatement or Paraphrasing	0.6	0.2
Pred.: Reflection of feelings	3.1	3.5
Pred.: Self-disclosure	2.1	16.2
Pred.: Affirmation and Reassurance	86.9	60.3
Pred.: Providing Suggestions	3.8	13.8
Pred.: Information	0.0	3.6
Pred.: Others	0.1	2.4

Section S6. Per-class metrics and calibration of the deterministic head

Across five training 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 expected calibration error 7.02 ± 3.37%. Per-class precision, recall, F1, and support are given in Table 5 and Figure 4 of the main text; the reliability diagram of a representative seed is shown in Figure S2.

Section S7. Embedding evaluation details

For each training seed, the 13 hyperbolic hierarchy embeddings learned inside CH-IQL were compared with the strategy tree: the Spearman correlation between pairwise geodesic distances and tree (path-length) distances over all 78 node pairs, and a scale-invariant distortion between the two distance sets. The Euclidean control (settings in Table S2)

directly optimizes a tree-distance stress objective and is therefore an upper reference rather than a like-for-like ablation.

Table S9. Structure preservation of the hierarchy embeddings.

Representation	Spearman ρ (p)	Distortion
Hyperbolic, seed 42	0.447 (4.1×10^{-5})	0.443
Hyperbolic, seed 123	−0.085 (0.46)	0.516
Hyperbolic, seed 2024	0.412 (1.8×10^{-4})	0.457
Hyperbolic, seed 3407	0.532 (5.4×10^{-7})	0.457
Hyperbolic, seed 7	0.240 (0.034)	0.474
Hyperbolic, mean ± SD	0.31 ± 0.22	0.47 ± 0.03
Euclidean control (64-d)	0.965	0.086

The Euclidean control satisfies all 12 hierarchy constraints.