BRIDGING THE SUPPORT-STRATEGY GROUNDING GAP: A THEORY-GUIDED FRAMEWORK FOR EMOTIONAL SUPPORT CONVERSATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotional support conversation (ESC) helps individuals cope with psychological distress through dialogue that provides understanding, comfort, and guidance. Effective machine support requires selecting a strategy that is contextually admissible and faithfully executing its intended function in the final response. Our empirical analysis shows that these two requirements can fail independently: a model may faithfully execute a contextually inadmissible strategy or fail to faithfully execute an admissible one. We call this the *support-strategy grounding gap*. We derive an information-theoretic lower bound that relates grounding success to contextual strategy admissibility and to the conditional information that the response conveys about the complete declared strategy beyond the dialogue context. Guided by this characterization, GROUNDESC first establishes explicit strategy-guided generation through structured supervised fine-tuning and then applies reinforcement learning signals for strategy selection, faithful execution, and response quality. Across multiple backbones, GROUNDESC consistently improves strategy selection and execution, substantially mitigating the grounding gap while maintaining response quality.

1 INTRODUCTION

Emotional support conversation (ESC) is a form of interpersonal communication that aims to alleviate distress through understanding, comfort, and encouragement as individuals disclose and work through difficult experiences (Beebe et al., 2011). It plays an important role in therapy, counseling, peer-support programs, and online mental-health services, where responsive communication can shape how individuals understand and cope with emotional difficulties (Sutton & Stewart, 2008; Mead, 2014; Riva, 2004). Unlike communication aimed at exchanging information or completing tasks, ESC centers on the seeker's emotional experience and evolving support needs.

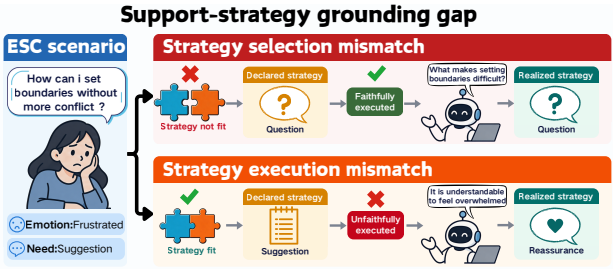


Figure 1: A declared strategy may mismatch the dialogue or fail to be faithfully executed in the response.

Helping Skills Theory provides a framework for adapting support to changing needs by distinguishing skills for facilitating exploration, insight, and action (Hill & O'Brien, 1999). Building on this framework, ESC research operationalizes these helping actions as support strategies that guide the supporter's behavior (Liu et al., 2021). According to Cutrona (1990), support effectiveness depends on how well a strategy's function matches the recipient's stressor and current needs. Comforting communication research (Jones & Wirtz, 2006; High & Dillard, 2012) further indicates that preserving a strategy's supportive function requires executing it through language that engages with the recipient's feelings and perspective. In ESC systems that explicitly declare a focal strategy before generating the final response, the strategy forms an intermediate interface between understanding the seeker's needs and responding to them. However, explicit declaration alone ensures

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 2: Strategy selection and response consistency across six models.

neither appropriate selection nor faithful execution. A system may faithfully execute a contextually inappropriate strategy, or select an appropriate strategy but fail to carry its intended supportive function into the response. As shown in Figure 1, we refer to either failure as the ***support-strategy grounding gap***. This grounding gap defines the central problem studied in this work: how to make an explicit support strategy both justified by the dialogue context and faithfully executed.

To examine the support-strategy grounding gap in practice, we assess whether selected strategies are contextually admissible and whether responses align with the declared strategies. For the latter, we use blind plan recovery (BPR), which predicts the focal strategy category from the dialogue history and final response, providing a category-level proxy for strategy execution. Figure 2 shows that the six models repeatedly fail to align these two aspects: some selected strategies are contextually inadmissible yet faithfully executed, whereas others are admissible but poorly executed. Thus, the gap is not tied to a particular model or a single failure mode. It reflects a recurring failure to jointly ground strategy selection and execution, motivating separate objectives for the two components.

These observations suggest that support-strategy grounding is governed by two complementary factors. We therefore develop an information-theoretic characterization of grounding success. We characterize strategy selection by p_{sel}, the probability that the selected strategy is contextually admissible, and strategy execution by I^{+}_{exe}, the information the final response conveys about the declared strategy beyond the context, conditional on successful selection. We define *grounding success* as the joint occurrence of contextual admissibility and faithful execution, and derive a lower bound on its probability. The bound increases with either p_{sel} or I^{+}_{exe} when the other quantity is fixed, showing why improving only one side cannot fully close the support-strategy grounding gap.

Guided by this characterization, we propose GROUNDESC (**Ground**ing Support Strategies in **E**motional **S**upport **C**onversation), a two-stage framework designed to close the support-strategy grounding gap. It first uses structured supervised fine-tuning to initialize explicit strategy-guided generation and then applies Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to refine strategy selection and execution. The selection objective rewards contextually admissible strategies, while the execution objective, motivated by I^{+}_{exe}, encourages faithful execution of the declared strategy. Complementing these grounding objectives, GROUNDESC uses a learned quality reward to evaluate contextual relevance, empathic attunement, support effectiveness, and naturalness. Experiments across four backbone models show that GROUNDESC consistently improves both sides of grounding while maintaining response quality.

2 THEORY OF SUPPORT-STRATEGY GROUNDING

The grounding gap identified in Figure 1 arises from two distinct failures: the declared strategy may be inappropriate for the seeker's current support needs, or an appropriate strategy may not be faith-

fully executed in the response. We characterize these two failure modes through *strategy selection* and *strategy execution*, which together determine support-strategy grounding. This analysis directly motivates the two grounding objectives introduced in Section 3.

Grounding as Selection and Execution. Let H denote the dialogue history, C the generated seeker-state analysis, R the final supportive response, and $X = (H, C)$. Let T denote the textual Strategy Line. For theoretical analysis, we associate T with its semantic strategy intent through a deterministic abstraction $\phi_x : \mathcal{T} \to \mathcal{Z}_x$ and define $Z = \phi_X(T)$. Here, Z retains the focal strategy category and its intended supportive function while abstracting away surface-level linguistic variation. The mapping ϕ_x is analytical and need not be explicitly computed during training.

The model generates the textual trajectory (C, T, R), whereas the analysis operates on its intent-level abstraction (C, Z, R). Let $\mathcal{A}$ denote the finite set of strategy categories, and let $a(z) \in \mathcal{A}$ extract the focal category from strategy intent z.

Definition 1 (Support-Strategy Grounding)**.** *For each context x, let $\mathcal{A}^+(x) \subseteq \mathcal{A}$ denote the nonempty set of contextually admissible strategy categories. Strategy selection succeeds when $S := \mathbf{1}\{a(Z) \in \mathcal{A}^+(X)\}, p_{\mathrm{sel}} := \Pr(S = 1) > 0$. Conditional on $S = 1$, let E denote successful MAP recovery of the declared strategy intent Z from (X, R). Successful support-strategy grounding is $\mathcal{G} := \{S = 1\} \cap E$, with corresponding probability $\Pr(\mathcal{G}) = p_{\mathrm{sel}} \Pr(E \mid S = 1)$.*

This decomposition mirrors the two observed grounding failures: S measures whether the declared strategy fits the current support needs, whereas E measures whether its intended function remains recoverable from the generated response after successful selection.

Why Grounding Requires Both Selection and Execution. We characterize the execution side of the grounding gap by relating faithful execution to uncertainty about the declared strategy intent.

Lemma 1 (MAP Recovery Inequality)**.** *Conditional on $S = 1$, let $f^\star(x, r) \in \arg\max_{z \in \mathrm{supp}(Z \mid X=x, S=1)} \Pr(Z = z \mid X = x, R = r, S = 1)$ be the MAP estimator of the declared strategy intent, and let $\widehat{Z} = f^\star(X, R)$, such that $E = \{\widehat{Z} = Z\}$. Then*

$$\Pr(E \mid S = 1) \geq \exp(-H(Z \mid X, R, S = 1)).$$

Lemma 1 relates execution success to the residual uncertainty about the declared strategy intent. However, $H(Z \mid X, R, S = 1)$ alone does not distinguish information already available from the context from information newly contributed by the response. Define the conditional strategy execution information as

$$I^+_{\mathrm{exe}} := \mathrm{I}(Z; R \mid X, S = 1) = \mathrm{H}(Z \mid X, S = 1) - \mathrm{H}(Z \mid X, R, S = 1),$$

which measures the information contributed by the response beyond the context. Assume that, for every relevant x, $|\mathrm{supp}(Z \mid X = x, S = 1)| \leq K_+$ for every relevant context x, where K_+ bounds the effective number of semantic strategy intents.

Theorem 1 (Grounding Achievability Bound)**.** *Under the bounded-support condition, the probability of successful support-strategy grounding satisfies*

$$\Pr(\mathcal{G}) \geq \frac{p_{\mathrm{sel}}}{K_+} \exp\big(I^+_{\mathrm{exe}}\big).$$

Theorem 1 formalizes the two failure modes observed in Figure 1: grounding depends jointly on selecting an admissible strategy and preserving its intent in the response. Improving execution cannot eliminate failures caused by inappropriate selection, while selection alone does not guarantee faithful execution. This motivates separate optimization signals for the two components. The execution side raises a further question: once an admissible strategy has been selected, what must the response preserve for that strategy to remain faithfully realizable? Let $P^+_{\mathrm{e}} := \Pr(\widehat{Z} \neq Z \mid S = 1) = 1 - \Pr(E \mid S = 1)$ denote the conditional execution error.

Theorem 2 (Execution-Information Converse)**.** *Suppose the bounded-support condition holds. If the conditional execution error satisfies $P^+_{\mathrm{e}} \leq \delta$ and $0 \leq \delta \leq 1 - \frac{1}{K_+}$, then necessarily*

$$I^+_{\mathrm{exe}} \geq H(Z \mid X, S = 1) - h(\delta) - \delta \log(K_+ - 1),$$

where $h(\delta) := -\delta \log \delta - (1 - \delta) \log(1 - \delta)$ is the binary entropy function.

Theorem 2 shows that reliable execution cannot be achieved with arbitrarily little execution information. As the allowable execution error decreases, the response must resolve more of the strategy uncertainty remaining after the context. Thus, for a prescribed execution reliability, faithful execution imposes a lower bound on I^+_{exe}.

Combining this converse with $\Pr(\mathcal{G}) = p_{\mathrm{sel}}(1-P^+_{\mathrm{e}})$ makes the asymmetry between the two grounding failures explicit: $\Pr(\mathcal{G}) \le p_{\mathrm{sel}}$, whereas reliable execution requires the response to resolve a sufficient portion of the strategy uncertainty that remains after selection.

From Execution Information to Observer Gain. The converse identifies I^+_{exe} as necessary for faithful strategy execution, but this actor-induced mutual information is not directly available during training. We therefore connect it to the prior-to-posterior information gain of a fixed strategy observer. Let $q_\tau(z \mid x)$ and $q_\tau(z \mid x, r)$ denote the observer's prior and posterior beliefs about the declared strategy intent. Define

$$\Delta_\tau(x, z, r) := \log q_\tau(z \mid x, r) - \log q_\tau(z \mid x), \quad I^+_\tau := \mathbb{E}_{(X,Z,R)\sim p_{\pi_\theta}(\cdot|S=1)}[\Delta_\tau(X, Z, R)].$$

Conditioning on $S = 1$ specifies the evaluation distribution; S is not supplied as an observer input.

Proposition 1 (Observer Approximation of Execution Information)**.** *The conditional observer information gain satisfies*

$$I^+_\tau = I^+_{\mathrm{exe}} + \epsilon^+_{\mathrm{pri}} - \epsilon^+_{\mathrm{post}},$$

where $\epsilon^+_{\mathrm{pri}}$ and $\epsilon^+_{\mathrm{post}}$ are the observer's expected KL approximation errors for the context-only and context-response conditional distributions, respectively. Consequently,

$$|I^+_\tau - I^+_{\mathrm{exe}}| \le \epsilon^+_{\mathrm{pri}} + \epsilon^+_{\mathrm{post}}.$$

Proposition 1 bridges theory and optimization: a fixed observer can approximate the execution quantity without direct access to actor-induced semantic distributions. Section 3 instantiates this principle at the textual Strategy-Line level. Together, the analysis yields two design requirements for GROUNDESC: promote contextually admissible strategy selection to increase p_{sel}, and observer-supported strategy execution to increase the execution signal associated with I^+_{exe}.

3 GROUNDESC FRAMEWORK

The analysis in Section 2 identifies strategy selection and execution as two asymmetric bottlenecks of support-strategy grounding and provides a theoretical bridge from ideal execution information to a practical training signal. Guided by this analysis, we develop GROUNDESC (**Ground**ing Support Strategies in **E**motional **S**upport **C**onversation), a two-stage framework that explicitly targets both factors. As illustrated in Figure 3, structured supervised fine-tuning first establishes an explicit strategy-guided generation process over seeker-state analysis, Strategy Line, and supportive response; reinforcement learning then improves contextually admissible strategy selection and faithful execution of the declared strategy intent while preserving response quality.

3.1 STRUCTURED SUPPORT INITIALIZATION

The first stage establishes the structured generation interface required for support-strategy grounding. Let $\mathcal{D}_{\mathrm{SFT}}$ contain training instances (h, c, t, r), where h is the dialogue history, c is the seeker-state analysis, t is the textual Strategy Line, and r is the final supportive response. We train the model autoregressively by minimizing

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(h,c,t,r)\sim\mathcal{D}_{\mathrm{SFT}}}\left[\log \pi_\theta(c, t, r \mid h)\right].$$

This structured objective equips the model with three capabilities: (1) **seeker-state analysis**, which identifies the situation, affective state, and support needs; (2) **explicit strategy formulation**, which produces a Strategy Line specifying a focal strategy and its intended function; and (3) **strategy-guided response generation**, which executes the declared strategy in natural language.

The resulting checkpoint initializes three models for reinforcement learning: a trainable actor π_θ for trajectory generation, a frozen reference policy π_{ref} for KL regularization, and a frozen strategy execution scorer π_τ that evaluates how strongly the response expresses the declared strategy.

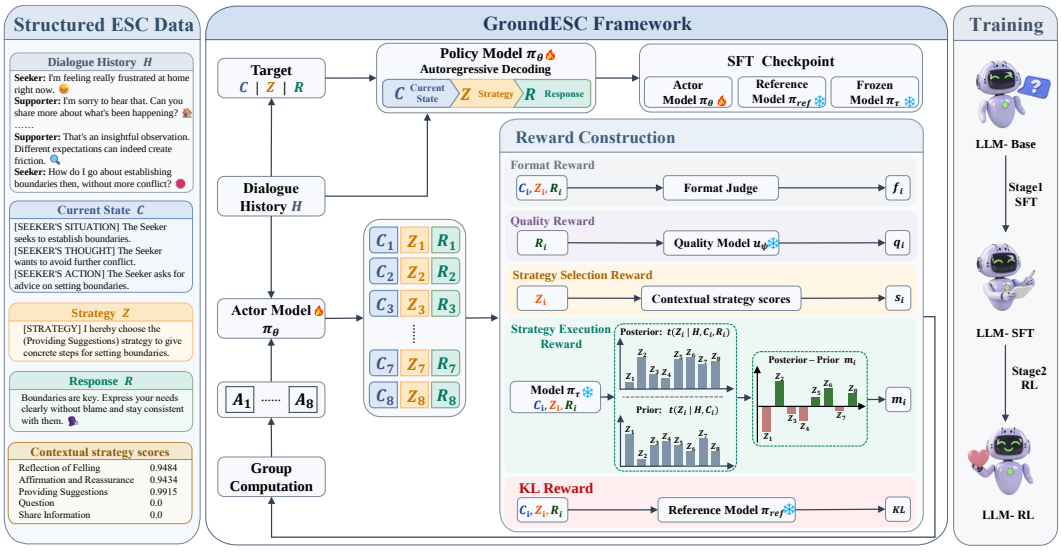


Figure 3: **Overview of GROUNDESC.**

3.2 SUPPORT-STRATEGY GROUNDING OBJECTIVES

We next instantiate the two grounding factors identified in Section 2 as trajectory-level optimization signals: contextual strategy selection and faithful strategy execution.

Contextual strategy selection. Because a dialogue context may admit multiple appropriate strategy categories, we evaluate the generated category against an admissible set rather than a single reference. For each context x, let $\mathcal{A}^+(x) \subseteq \mathcal{A}$ denote its admissible strategy categories, and let $\omega_x(a) > 0$ denote the soft suitability score assigned to each $a \in \mathcal{A}^+(x)$.

For a generated textual trajectory (c_i, t_i, r_i), let $x_i = (h, c_i)$ and $z_i = \phi_{x_i}(t_i)$, where z_i is the semantic strategy intent represented by the generated Strategy Line t_i. Because t_i explicitly specifies its focal strategy category, we extract $a(z_i)$ and define the category-level selection reward as

$$s_i = \begin{cases} \omega_{x_i}(a(z_i)), & a(z_i) \in \mathcal{A}^+(x_i), \\ 0, & a(z_i) \notin \mathcal{A}^+(x_i). \end{cases}$$

Thus, s_i assigns a positive suitability score to an admissible category and zero otherwise. In particular, $\mathbf{1}\{s_i > 0\}$ is the trajectory-level counterpart of the selection event S. The reward s_i therefore directly promotes the context-to-strategy component characterized by p_{sel} in Theorems 1-2.

Faithful strategy execution. In contrast to the category-level selection objective, strategy execution concerns the complete declared intent Z, including both its focal category and intended supportive function. As established in Section 2, execution among successfully selected trajectories is characterized by $I^+_{\text{exe}} = \mathrm{I}(Z; R \mid X, S = 1)$. This quantity measures how much additional information the response provides about the declared strategy intent beyond what is already available from the context, conditional on the selected strategy category being contextually admissible.

Evaluating I^+_{exe} during policy optimization is impractical because it depends on the actor-induced conditional distributions over semantic strategy intents. Proposition 1 connects this ideal quantity to the prior-to-posterior information gain of a fixed strategy observer, up to its approximation errors. We instantiate this principle at the textual level using the frozen strategy execution scorer π_τ.

Since the semantic abstraction $Z = \phi_X(T)$ is introduced for analysis rather than explicitly computed during optimization, we instantiate the observer gain directly on the generated textual Strategy Line T. Specifically, for the generated Strategy Line $t_i = (t_{i,1}, \ldots, t_{i,L_i})$, the scorer evaluates the same textual strategy once from the context alone and once with the generated response included. We define the length-normalized textual observer gain as

$$m_i = \frac{1}{L_i} \sum_{\ell=1}^{L_i} \Big[\underbrace{\log \pi_\tau \left(t_{i,\ell} \mid x_i, r_i, t_{i,<\ell}\right)}_{\text{posterior strategy score}} - \underbrace{\log \pi_\tau \left(t_{i,\ell} \mid x_i, t_{i,<\ell}\right)}_{\text{prior strategy score}} \Big]$$

Here, $\log \pi_\tau(\cdot)$ denotes the token log probability computed under teacher forcing, and division by L_i converts each sequence score into an average over the Strategy Line. The score m_i is a computable text-level surrogate for the semantic observer information gain, rather than an unbiased per-trajectory of I^+_{exe}. A positive m_i indicates that observing the response increases the frozen scorer's support for the declared Strategy Line beyond what is already predictable from the context.

Because the scale of m_i can vary across focal strategy categories, we apply category-specific calibration. For each category $k \in \mathcal{A}$, let b_k and $u_k > b_k$ denote its calibration center and upper reference, respectively. We define the calibrated execution score as

$$e_i = \text{clip}\left(\frac{m_i - b_{a(z_i)}}{u_{a(z_i)} - b_{a(z_i)}}, 0, 1\right).$$

The focal category $a(z_i)$ determines the calibration parameters. This monotone transformation normalizes category-dependent scales while preserving m_i ordering within each category.

Finally, we restrict the execution reward to trajectories whose declared strategy category is contextually admissible $E_i = \mathbf{1}\{s_i > 0\}\, e_i$. This gating is the sample-level counterpart of conditioning execution on $S = 1$ in Definition 1: execution evidence contributes to policy optimization only when the declared strategy category is contextually admissible. Importantly, the selection event is not provided as an additional input to the frozen scorer π_τ; it determines only whether the resulting execution reward is activated. Together, s_i and E_i instantiate the two grounding components identified by Theorems 1-2: s_i promotes context-to-strategy appropriateness, while E_i promotes faithful strategy-to-response execution after an admissible strategy has been selected.

3.3 Quality-Aware Group Optimization

Supportive response quality. Optimizing strategy selection and execution alone may encourage responses with highly recognizable strategy-specific patterns. Such responses can select an appropriate strategy and express its intended function clearly, yet still become formulaic, mechanically phrased, or overly tied to stereotyped executions of a strategy. This may limit the diversity and naturalness of emotional support, even when support-strategy grounding itself is successful. We therefore complement the grounding objectives with an explicit response-quality signal.

Specifically, we train a quality reward model u_ψ along four dimensions: **contextual relevance**, whether the response addresses the seeker's current situation and support needs; **empathic attunement**, whether it appropriately responds to the seeker's emotional state; **support effectiveness**, whether it provides helpful and suitable support; and **naturalness**, whether it is fluent, specific, and free from rigid or repetitive expressions. The resulting reward is $q_i = u_\psi(h, r_i) \in [0, 1]$.

Composite reward. We additionally use a rule-based format reward $f_i = \mathcal{F}(c_i, t_i, r_i) \in [0, 1]$ to encourage the required analysis–Strategy Line–response structure. We denote the four reward components as $\mathsf{F}_i = f_i$ (format), $\mathsf{S}_i = s_i$ (strategy selection), $\mathsf{E}_i = \mathbf{1}\{s_i > 0\}e_i$ (strategy execution), and $\mathsf{Q}_i = q_i$ (response quality), where the indicator restricts execution rewards to contextually admissible strategies. The total trajectory reward is

$$\mathcal{R}_i = \lambda_f \mathsf{F}_i + \lambda_s \mathsf{S}_i + \lambda_e \mathsf{E}_i + \lambda_q \mathsf{Q}_i, \quad \lambda_f, \lambda_s, \lambda_e, \lambda_q \geq 0$$

Here, S_i and E_i instantiate the selection and execution objectives from the theoretical analysis, while F_i enforces structured generation and Q_i preserves overall response quality.

Group policy optimization. For each dialogue history h, we sample a group of G trajectories $y_i = (c_i, t_i, r_i) \sim \pi_{\theta_{\text{old}}}(\cdot \mid h)$. Following group-relative policy optimization, we normalize the trajectory rewards within each group:

$$\widehat{A}_i = \frac{\mathcal{R}_i - \mu_{\mathcal{R}}}{\sigma_{\mathcal{R}} + \delta},$$

where μ_R and σ_R are the mean and standard deviation of the rewards within the sampled group, and $\delta > 0$ is a small constant for numerical stability. The actor is then optimized with the standard clipped GRPO objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{\ell=1}^{|y_i|}\min\left(\rho_{i,\ell}\widehat{A}_i,\ \text{clip}(\rho_{i,\ell}, 1-\varepsilon, 1+\varepsilon)\widehat{A}_i\right)\right] + \beta\mathcal{L}_{\text{KL}},$$

Table 1: Overall performance on the SSConv test set.

Model	Adm. (%)↑	GSF↑	CR↑	EA↑	SE↑	Nat.↑
Specialized ESC Systems						
Kardia-R1 Yuan et al. (2026)	-	-	3.037	2.979	2.475	2.755
Psyche-R1 Dai et al. (2026)	54.77	2.018	2.799	2.667	2.474	2.739
PsyLLM-8B Hu et al. (2025)	55.31	2.024	3.641	3.613	3.413	3.412
Prompted Foundation Models						
Llama3.1-8B-Instruct (Grattafiori et al., 2024)	45.09	1.858	3.487	3.209	2.887	3.347
Llama3.1-70B-Instruct (Grattafiori et al., 2024)	50.64	2.252	3.625	3.377	3.194	3.549
Qwen2.5-14B-Instruct (Hui et al., 2024)	63.92	2.628	3.696	3.185	3.024	3.429
Qwen3-8B-Instruct (Yang et al., 2025)	60.00	2.513	3.430	3.200	2.884	3.162
Qwen3-14B-Instruct (Yang et al., 2025)	68.23	2.741	3.729	3.349	3.126	3.377
Qwen3-32B-Instruct (Yang et al., 2025)	64.55	2.839	3.937	3.620	3.346	3.465
Kimi-K2-Instruct (Team et al., 2025)	60.35	2.804	**4.15**	**3.875**	<u>3.637</u>	3.865
GLM-4.5 (Zeng et al., 2025)	62.49	2.637	4.00	3.771	3.533	3.556
DeepSeek-R1 (Guo et al., 2025)	56.56	2.431	3.958	3.701	3.589	3.636
DeepSeek-V3 (Liu et al., 2024)	70.45	2.970	<u>4.026</u>	<u>3.826</u>	3.316	3.739
GROUNDESC Models						
Qwen2.5-1.5B-Instruct	71.48	2.669	3.824	3.328	3.463	3.751
Qwen2.5-3B-Instruct	71.38	<u>3.065</u>	3.875	3.406	3.616	<u>3.889</u>
Qwen2.5-7B-Instruct	**75.13**	**3.294**	3.926	3.610	**3.724**	**3.987**
Gemma-2B-Instruct	<u>72.51</u>	2.993	3.839	3.330	3.354	3.659

where $\rho_{i,\ell} = \pi_\theta(y_{i,\ell} \mid h, y_{i,<\ell}) / \pi_{\theta_{\text{old}}}(y_{i,\ell} \mid h, y_{i,<\ell})$, and $\mathcal{L}_{\text{KL}}$ regularizes the actor toward the frozen reference policy π_{ref}. Through this objective, GROUNDESC jointly optimizes appropriate strategy selection, faithful strategy execution, and response quality while preserving the structured generation behavior established during supervised fine-tuning.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. All training and evaluation data are derived from SSConv Chen et al. (2025). Structured supervised fine-tuning uses $19{,}849$ quality-controlled examples containing a seeker-state analysis, an explicit Strategy Line, and the original supporter response. For reinforcement learning, we construct $7{,}883$ dialogue contexts, each annotated with one to three admissible strategies and soft appropriateness scores to support multiple valid decisions. We further construct $6{,}000$ response-preference pairs to train the quality reward model on contextual relevance, empathic attunement, support effectiveness, and naturalness. Further details are provided in Appendix C.2.

Evaluation metrics. We organize evaluation into two categories: Support-Strategy Grounding and Response Quality. (1) **Support-Strategy Grounding** assesses whether the model selects an appropriate strategy and faithfully executes it in the response. Let z_i denote the strategy declared for response i, $\mathcal{A}_i$ the set of admissible strategies, and $\text{SF}_i \in [1, 5]$ its execution fidelity. We report strategy admissibility (Adm) and Gated Strategy Fidelity (GSF) as $\text{GSF} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[z_i \in \mathcal{A}_i]\,\text{SF}_i$ and $\text{Adm} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[z_i \in \mathcal{A}_i]$. Adm measures appropriate strategy selection, while GSF jointly captures selection and execution, assigning zero credit to inadmissible, invalid, or missing declarations. (2) **Response Quality** evaluates contextual relevance (CR), empathic attunement (EA), support effectiveness (SE), and naturalness (Nat.). Evaluation is conducted on $2{,}714$ SSConv supporter turns using DeepSeek-V4-Flash (Xu et al., 2026), which is distinct from the training reward model. Full scoring rubrics, evaluator prompts, and aggregation procedures are provided in Appendix C.3.

4.2 MAIN RESULTS

Table 1 compares GROUNDESC with specialized ESC systems and prompted foundation models. Among specialized systems with reported grounding scores, the best admissibility and GSF are 55.31% and 2.024, respectively, while DeepSeek-V3, the strongest prompted baseline, reaches 70.45% admissibility and a GSF of 2.970. All four GROUNDESC variants achieve admissibility rates of 71.38–75.13%, exceeding every baseline with reported grounding scores. The 3B and 7B

Table 2: Component ablations. F, Q, S, E denote format, quality, selection, and execution rewards.

Backbone	Training	Reward				Support-Strategy Grounding		Response Quality			
		F	Q	S	E	Adm. (%)↑	GSF↑	CR↑	EA↑	SE↑	Nat.↑
Qwen2.5-3B-Instruct	SFT	–	–	–	–	52.25	2.153	3.563	3.307	3.312	3.495
	RL	✓	✓	–	–	51.28	2.155	3.812	3.250	3.530	3.576
	RL	✓	✓	✓	–	68.67	2.516	**3.912**	3.273	3.592	3.699
	RL	✓	✓	✓	✓	**71.38**	**3.065**	3.875	**3.406**	**3.616**	**3.889**
Gemma-2B-Instruct	SFT	–	–	–	–	52.91	2.144	3.364	2.888	3.035	3.092
	RL	✓	✓	–	–	52.67	2.189	3.634	3.058	3.200	3.496
	RL	✓	✓	✓	–	69.89	2.521	3.678	3.196	3.228	3.447
	RL	✓	✓	✓	✓	**72.51**	**2.993**	**3.839**	**3.330**	**3.354**	**3.659**

Table 3: Ablation of execution-score formulations m_i on two backbones.

Backbone	Execution Signal	Adm. (%)↑	GSF↑	CR↑	EA↑	SE↑	Nat.↑
Qwen2.5-3B-Instruct	Prior only	71.33	2.738	3.851	3.281	3.410	3.596
	Posterior only	71.29	2.863	3.822	3.276	3.378	3.548
	Posterior–Prior (Ours)	71.38	3.065	3.875	3.406	3.616	3.889
Gemma-2B-Instruct	Prior only	73.32	2.650	3.711	3.196	3.181	3.443
	Posterior only	72.69	2.730	3.676	3.185	3.178	3.461
	Posterior–Prior (Ours)	72.51	2.993	3.839	3.330	3.354	3.659

variants surpass DeepSeek-V3 in GSF, reaching 3.065 and 3.294, with the 7B model obtaining the best overall results on both grounding metrics. This advantage is not confined to larger backbones, as the 1.5B and Gemma-2B variants also outperform all prompted models in admissibility and specialized ESC systems in GSF. Taken together, these results show that our optimization improves appropriate strategy selection and faithful strategy execution across architectures and scales.

For response quality, the substantially larger Kimi-K2-Instruct and DeepSeek-V3 models remain ahead in contextual relevance and empathic attunement. Nevertheless, GROUNDESC-7B achieves the best support effectiveness (3.724) and naturalness (3.987), while the 3B variant ranks second in naturalness (3.889). This profile reflects our objective design: response-quality rewards preserve quality, while optimization improves support-strategy grounding, enabling compact backbones to achieve the strongest grounding performance with leading support effectiveness and naturalness.

Table 2 demonstrates the complementary effects of the training components. Format and quality rewards improve response quality while largely preserving grounding performance. Adding the strategy-selection reward increases admissibility by over 16 percentage points on both backbones, together with clear GSF gains. The execution reward further raises GSF from 2.516 to 3.065 on Qwen2.5-3B and from 2.521 to 2.993 on Gemma-2B, with comparatively smaller changes in admissibility. These results show that selection mainly improves contextual strategy choice, while execution strengthens faithful execution, enabling the full objective to achieve the strongest grounding performance without sacrificing response quality.

Table 5: Comparison of GROUNDESC with and without E_i. $\mathrm{SF_H}$: $\mathrm{SF_{Human}}$. $\mathrm{SF_L}$: $\mathrm{SF_{LLM}}$.

Variant	$\mathrm{SF_H}$	$\mathrm{SF_L}$	L_{prior}	L_{post}	m_i
w/o E_i	3.368	3.500	-0.457	-0.446	0.012
Full	4.022	4.000	-0.465	-0.402	0.063
Δ	**+0.653**	**+0.500**	-0.007	**+0.044**	**+0.052**

Table 6: Association between frozen-observer scores and SF.

m_i formulation	$\rho_s(\mathrm{SF})$ (95% CI)
L_{prior}	-0.086 [-0.176, 0.007]
L_{post}	0.181 [0.082, 0.276]
$L_{\mathrm{post}} - L_{\mathrm{prior}}$	**0.349** [0.254, 0.437]

Tables 3, 5, and 6 jointly examine the design of the strategy execution reward. Table 3 shows that the posterior–prior formulation consistently achieves higher GSF than using either the prior or posterior alone on both backbones, while admissibility remains similar. This supports the role of the difference term: the prior mainly reflects strategy predictability from context, whereas the posterior alone can conflate this contextual information with evidence contributed by the response; their difference better isolates the response-induced execution signal. Tables 5 and 6 provide complementary validation. Including E_i substantially improves human-rated Strategy Fidelity and increases the raw observer gain. The improvement is driven mainly by stronger posterior evidence, with only a minor change in the prior, supporting the intended role of E_i in strengthening response-induced strategy execution. Consistently, the posterior–prior score shows the strongest association with hu-

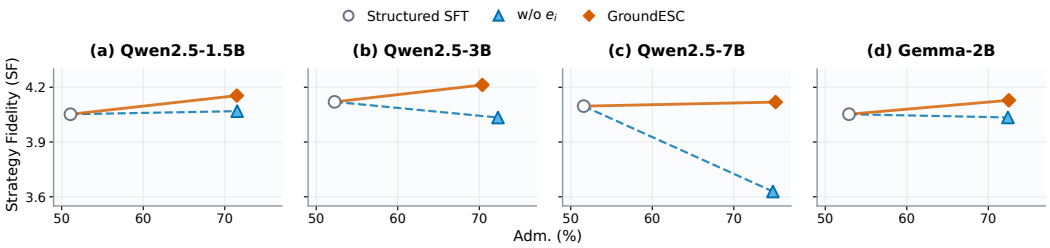


Figure 4: Support-strategy grounding across four backbones. Strategy admissibility and fidelity are compared for Structured SFT, GROUNDESC without E_i, and full GROUNDESC.

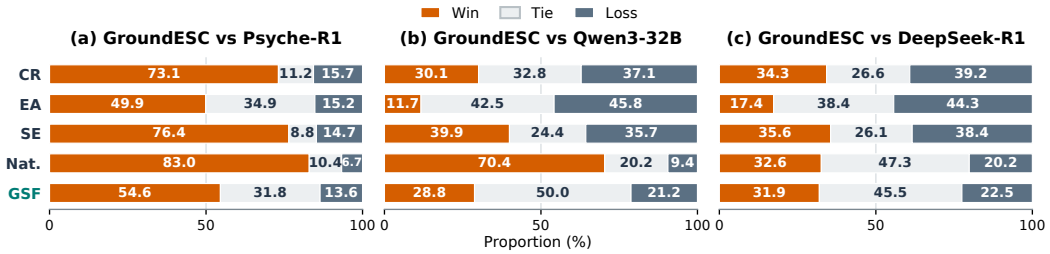


Figure 5: A/B evaluation of GROUNDESC(Qwen2.5-3B) against three representative systems.

Table 4: Zero-shot transfer to the ESConv test set after training only on SSConv.

Backbone	Variant	Adm. (%)↑	GSF↑	CR↑	EA↑	SE↑	Nat.↑
Qwen2.5-7B	Structured SFT	38.50	1.510	3.275	2.800	2.560	3.015
	GROUNDESC	**47.50**	**1.855**	**3.465**	**3.045**	**2.860**	**3.320**
Gemma-2B	Structured SFT	39.50	1.545	3.185	2.700	2.570	3.020
	GROUNDESC	**46.50**	**1.815**	**3.350**	**2.890**	**2.790**	**3.320**

man fidelity ($\rho_s = 0.349$), outperforming either component alone. Together, these results support the posterior–prior difference as an empirical proxy for the strategy execution signal characterized by I_{exe}. Figure 4 shows that optimization without E_i mainly improves admissibility, while adding E_i increases strategy fidelity without sacrificing admissibility. Therefore, GROUNDESC improves strategy selection and execution across backbones, rather than either aspect.

To complement the pointwise evaluation, we conduct A/B comparisons between GROUNDESC (Qwen2.5-3B) and three representative systems, using DeepSeek-V4-Flash as the evaluator. As shown in Figure 5, GROUNDESC (Qwen2.5-3B) achieves more GSF wins than losses against all three baselines, including 54.6% vs. 13.6% against Psyche-R1. It also shows advantages in naturalness while remaining competitive in support effectiveness and contextual relevance. Overall, the pairwise results further support strong grounding and response quality with a compact 3B backbone.

Table 4 evaluates zero-shot transfer to ESConv after training only on SSConv. Across both backbones, GROUNDESC consistently outperforms Structured SFT on support-strategy grounding and all response-quality metrics. The consistent gains across grounding and response quality further suggest that the learned behavior is not specific to the SSConv training distribution. These results show that grounding improvements transfer beyond SSConv and generalize to another ESC dataset.

5 CONCLUSION

Reliable use of an explicit support strategy requires selecting a contextually admissible strategy and faithfully executing its intended function in the final response. We formalize this joint requirement as support-strategy grounding and derive an information-theoretic bound connecting grounding success to strategy selection and response informativeness about the declared intent. Guided by this analysis, GROUNDESC jointly optimizes strategy selection and execution, yielding consistent gains in admissibility and GSF across four backbones while preserving response quality and transferring to ESConv. Together, these results highlight the importance of explicitly aligning context, declared strategy, and response in machine ESC systems.

AI USE STATEMENT

AI tools assisted manuscript organization, mathematical checking, and editing. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Steven A Beebe, Susan J Beebe, and Mark V Redmond. *Interpersonal communication: Relating to others*. Pearson Education Boston, 2011.

Zhuang Chen, Yaru Cao, Guanqun Bi, Jincenzi Wu, Jinfeng Zhou, Xiyao Xiao, Si Chen, Hongning Wang, and Minlie Huang. Socialsim: Towards socialized simulation of emotional support conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1274–1282, 2025.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3014–3026, 2022.

Carolyn E Cutrona. Stress and social support—in search of optimal matching. *Journal of social and clinical Psychology*, 9(1):3–14, 1990.

Chongyuan Dai, Jinpeng Hu, Hongchang Shi, Zhuo Li, Dan Guo, Xun Yang, and Meng Wang. Psyche-r1: Towards reliable psychological llms through unified empathy, expertise, and reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24889–24906, 2026.

Mengyi Deng, Zhiwei Li, Xin Li, Tingyu Zhu, Ying Zhao, Zhijiang Guo, and Wei Wang. Uncertainty-aware clarification in llm agents with information gain. *arXiv preprint arXiv:2606.03135*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Andrew C High and James Price Dillard. A review and meta-analysis of person-centered messages and social support outcomes. *Communication Studies*, 63(1):99–118, 2012.

Clara E Hill and Karen M O'Brien. Helping skills. *Washington, DC: American Psychological Association*, 1999.

He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, Laizhong Cui, and Qi Tian. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *arXiv preprint arXiv:2505.15715*, 2025.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Susanne M Jones and John G Wirtz. How does the comforting process work? an empirical test of an appraisal-based model of comforting. *Human Communication Research*, 32(3):217–243, 2006.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 110–119, 2016.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 3469–3483, 2021.

Shery Mead. *Intentional peer support: An alternative approach*, volume 1. Intentional Peer Support West Chesterfield, 2014.

Ashwin Paranjape and Christopher D Manning. Human-like informative conversations: Better acknowledgements using conditional mutual information. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 768–781, 2021.

Giuseppe Riva. Cybertherapy: Internet and virtual reality as assessment and rehabilitation tools for clinical psychology and neuroscience. *(No Title)*, 2004.

Jonathan Scarlett and Volkan Cevher. An introductory guide to fano's inequality with applications in statistical estimation. *arXiv preprint arXiv:1901.00555*, 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jan Sutton and William Stewart. *Learning to counsel: How to develop the skills, insight and knowledge to counsel others*. Hachette UK, 2008.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory. *arXiv preprint arXiv:2411.11984*, 2024.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 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)*, pp. 308–319, 2022.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. Bottlesum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. 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)*, pp. 3752–3761, 2019.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xixian Yong, Xiao Zhou, Yingying Zhang, Jinlin Li, Yefeng Zheng, and Xian Wu. Think or not? exploring thinking efficiency in large reasoning models via an information-theoretic lens. *Advances in Neural Information Processing Systems*, 38:2787–2827, 2026.

Jiahao Yuan, Zhiqing Cui, Hanqing Wang, Yuansheng Gao, Yucheng Zhou, and Usman Naseem. Kardia-r1: Unleashing llms to reason toward understanding and empathy for emotional support via rubric-as-judge reinforcement learning. In *Proceedings of the ACM Web Conference 2026*, pp. 9230–9240, 2026.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1714–1729, 2023.

A Related Work

A.1 Emotional Support Conversation

Emotional support conversation (ESC) was formalized as a distinct dialogue task by Liu et al. (2021), who introduced ESConv with support strategy annotations grounded in Helping Skills Theory. This formulation established support strategy as an explicit intermediate decision in emotional dialogue, separating the choice of a supportive action from the wording of the final response. SocialSim subsequently constructed SSConv through socially informed simulation, providing larger and more diverse training resources and serving as the primary data foundation for our study (Chen et al., 2025).

Early ESC systems mainly improved strategy-aware generation by enriching the representation of the seeker's mental state or by organizing multiple support strategies. MISC combines fine-grained mental state modeling with mixed strategy generation (Tu et al., 2022), while MultiESC uses lookahead planning to select strategies according to their anticipated effects on subsequent dialogueturns (Cheng et al., 2022). Supporter further introduces reinforcement learning to optimize longer-term emotional outcomes rather than only the quality of the current response (Zhou et al., 2023). These studies establish the importance of strategy-level decision making, but their evaluation generally treats strategy choice and response generation as separate aspects of system behavior.

Recent work with large language models has increasingly focused on specialized reasoning and optimization for emotional support and mental-health counseling. PsyLLM-8B integrates diagnostic and therapeutic reasoning into psychological support generation (Hu et al., 2025), while Psyche-R1 jointly emphasizes empathy, expertise, and reasoning for reliable psychological assistance (Dai et al., 2026). Kardia-R1 further improves emotional-support reasoning through rubric-guided reinforcement learning toward understanding and empathy (Yuan et al., 2026). More broadly, work on reasoning faithfulness shows that explicit intermediate text does not guarantee that its content determines the final output (Turpin et al., 2023). This is particularly relevant to ESC systems that declare a support strategy before generating a response, since a plausible declaration alone ensures neither contextual appropriateness nor faithful execution.

A.2 Information Theory for Language Models

Information theory provides principled tools for quantifying statistical dependence, uncertainty, and the information preserved when one representation is derived from another. These tools have been used to study how a system retains task-relevant information while discarding irrelevant variation. For example, the information bottleneck principle has been applied to unsupervised and self-supervised sentence summarization, where the objective is to compress the input while preserving information useful for the summary (West et al., 2019). This perspective is closely related to structured generation settings in which an intermediate representation is expected to retain information that should influence a downstream output.

In dialogue generation, mutual information objectives have been introduced to discourage generic responses and to encourage responses that better reflect their conditioning contexts (Li et al., 2016; Paranjape & Manning, 2021). These methods treat the relationship between the context and the response as an explicit modeling target, complementing likelihood objectives that may favor frequent but weakly informative responses. Their central insight is that generation quality depends not only on whether a response is fluent, but also on how specifically it is related to the information that conditions the conversation.

Recent studies extend information-theoretic analysis to large language models and multi-step reasoning. Information measures have been used to characterize how information is distributed across chain-of-thought steps and how such patterns relate to reasoning behavior (Ton et al., 2024). Entropy based criteria further provide signals for deciding whether additional reasoning is useful, enabling adaptive termination of lengthy reasoning processes (Yong et al., 2026). Information gain has also been used to train agents to request clarification when the user's intent remains uncertain (Deng et al., 2026). Together, these studies show that information quantities can serve both as diagnostic measures of intermediate processes and as training signals for controlling downstream generation.

B Proofs for the Information-Theoretic Analysis

This appendix provides detailed proofs for the results in Section 2. Unless otherwise specified, all probabilities, entropies, and mutual information quantities are induced by the policy π_θ together with the underlying dialogue-context distribution. All logarithms are natural.

B.1 Proof of Lemma 1

Proof. For any execution (x, r) with positive probability under $S = 1$, define $p_{x,r}(z) := \Pr(Z = z \mid X = x, R = r, S = 1)$. Because $\widehat{Z}$ is the MAP estimator,

$$\Pr(E \mid X = x, R = r, S = 1) = \max_z p_{x,r}(z).$$

For any discrete distribution p, let $p_{\max} = \max_z p(z)$. Since $p(z) \le p_{\max}$,

$$H(p) = \sum_z p(z) \log \frac{1}{p(z)} \ge \log \frac{1}{p_{\max}},$$

so $p_{\max} \ge e^{-H(p)}$. Hence

$$\Pr(E \mid X = x, R = r, S = 1) \ge e^{-H(Z|X=x,R=r,S=1)}.$$

Averaging over $(X, R) \mid S = 1$ and applying Jensen's inequality to the convex function e^{-u} gives

$$\Pr(E \mid S = 1) \ge e^{-H(Z|X,R,S=1)}.$$

□

B.2 Proof of Theorem 1

Proof. By Definition 1 and Lemma 1,

$$\Pr(\mathcal{G}) \ge p_{\text{sel}} e^{-H(Z|X,R,S=1)}.$$

By the definition of conditional strategy, i.e., $I^+_{\text{exe}} := I(Z; R \mid X, S = 1) = H(Z \mid X, S = 1) - H(Z \mid X, R, S = 1)$,

$$H(Z \mid X, R, S = 1) = H(Z \mid X, S = 1) - I^+_{\text{exe}}.$$

The support assumption implies $H(Z \mid X = x, S = 1) \le \log K_+$ for every relevant x, and therefore $H(Z \mid X, S = 1) \le \log K_+$. Thus

$$\Pr(\mathcal{G}) \ge p_{\text{sel}} e^{-\log K_+ + I^+_{\text{exe}}} = \frac{p_{\text{sel}}}{K_+} e^{I^+_{\text{exe}}}.$$

This completes the proof. □

Remark on the semantic support size. The support condition applies to the semantic intent Z, not the textual Strategy Line T. Multiple textual executions may map to the same semantic intent under ϕ_x; thus K_+ characterizes effective semantic ambiguity rather than lexical diversity.

B.3 Proof of Theorem 2

Proof. For each context x, define

$$P_{\text{e}}(x) := \Pr(\widehat{Z} \ne Z \mid X = x, S = 1).$$

Then

$$P^+_{\text{e}} = \mathbb{E}_{X|S=1}[P_{\text{e}}(X)].$$

For fixed x, $Z \mid X = x, S = 1$ has support size at most K_+. Fano's inequality (Scarlett & Cevher, 2019) gives

$$H(Z \mid X = x, R, S = 1) \le h(P_{\text{e}}(x)) + P_{\text{e}}(x) \log(K_+ - 1).$$

Averaging over $X \mid S = 1$ and using concavity of h yields

$$H(Z \mid X, R, S = 1) \le h(P_{\mathrm{e}}^{+}) + P_{\mathrm{e}}^{+} \log(K_{+} - 1).$$

Therefore

$$I_{\mathrm{exe}}^{+} \ge H(Z \mid X, S = 1) - h(P_{\mathrm{e}}^{+}) - P_{\mathrm{e}}^{+} \log(K_{+} - 1).$$

Define $g(u) = h(u) + u \log(K_{+} - 1)$. For $u \in (0, 1)$,

$$g'(u) = \log \frac{(K_{+} - 1)(1 - u)}{u},$$

so g is nondecreasing on $[0, 1 - 1/K_{+}]$. Since $P_{\mathrm{e}}^{+} \le \delta \le 1 - 1/K_{+}$, we have $g(P_{\mathrm{e}}^{+}) \le g(\delta)$, yielding

$$I_{\mathrm{exe}}^{+} \ge H(Z \mid X, S = 1) - h(\delta) - \delta \log(K_{+} - 1).$$

This completes the proof. □

Near-perfect execution. As $\delta \to 0$, both $h(\delta)$ and $\delta \log(K_{+} - 1)$ vanish. Since $I_{\mathrm{exe}}^{+} \le H(Z \mid X, S = 1)$, the zero-error limit satisfies

$$I_{\mathrm{exe}}^{+} \to H(Z \mid X, S = 1).$$

B.4 Necessary Conditions for Target Grounding

Corollary 1 (Necessary Conditions for Target Grounding)**.** *Suppose* $\Pr(\mathcal{G}) \ge \gamma$ *for some* $\gamma \in (0, 1]$. *Then necessarily*

$$p_{\mathrm{sel}} \ge \gamma.$$

Moreover, whenever

$$\frac{\gamma}{p_{\mathrm{sel}}} \ge \frac{1}{K_{+}},$$

the execution information must satisfy

$$I_{\mathrm{exe}}^{+} \ge H(Z \mid X, S = 1) - h\left(1 - \frac{\gamma}{p_{\mathrm{sel}}}\right) - \left(1 - \frac{\gamma}{p_{\mathrm{sel}}}\right) \log(K_{+} - 1).$$

Proof. By Definition 1,

$$\Pr(\mathcal{G}) = p_{\mathrm{sel}} \left(1 - P_{\mathrm{e}}^{+}\right).$$

Since $1 - P_{\mathrm{e}}^{+} \le 1$, we have $\Pr(\mathcal{G}) \le p_{\mathrm{sel}}$. Thus, $\Pr(\mathcal{G}) \ge \gamma$ necessarily implies $p_{\mathrm{sel}} \ge \gamma$.

Moreover,

$$P_{\mathrm{e}}^{+} \le 1 - \frac{\gamma}{p_{\mathrm{sel}}}.$$

If $\gamma / p_{\mathrm{sel}} \ge 1/K_{+}$, then

$$1 - \frac{\gamma}{p_{\mathrm{sel}}} \le 1 - \frac{1}{K_{+}}.$$

Applying Theorem 2 with

$$\delta = 1 - \frac{\gamma}{p_{\mathrm{sel}}}$$

gives

$$I_{\mathrm{exe}}^{+} \ge H(Z \mid X, S = 1) - h\left(1 - \frac{\gamma}{p_{\mathrm{sel}}}\right) - \left(1 - \frac{\gamma}{p_{\mathrm{sel}}}\right) \log(K_{+} - 1),$$

which proves the result. □

Corollary 1 makes the asymmetry between strategy selection and execution explicit. Inappropriate selection imposes the hard ceiling $\Pr(\mathcal{G}) \le p_{\mathrm{sel}}$, which improved execution alone cannot overcome. Once an admissible strategy is selected, achieving reliable grounding further requires the response to preserve sufficient information about the declared strategy intent. These two necessary conditions motivate the complementary selection and execution objectives in GroundESC.

B.5 Selection–Execution Necessary Boundary

For a fixed policy π, define

$$H_+(\pi) := H_\pi(Z \mid X, S=1), \qquad p_{\mathrm{sel}}(\pi) := \Pr_\pi(S=1).$$

For target grounding probability γ, define the policy-specific converse boundary

$$\mathcal{B}_\gamma(\pi) := H_+(\pi) - h\left(1 - \frac{\gamma}{p_{\mathrm{sel}}(\pi)}\right) - \left(1 - \frac{\gamma}{p_{\mathrm{sel}}(\pi)}\right)\log(K_+ - 1).$$

Whenever $\gamma \le p_{\mathrm{sel}}(\pi) \le \min\{1, K_+\gamma\}$, any policy satisfying $\Pr_\pi(\mathcal{G}) \ge \gamma$ must obey

$$I^+_{\mathrm{exe}}(\pi) \ge \mathcal{B}_\gamma(\pi).$$

This is a necessary, policy-specific converse boundary rather than a universal attainable frontier determined by p_{sel} alone, because $H_+(\pi)$ may itself vary across policies.

B.6 Proof of Proposition 1

Define

$$\epsilon^+_{\mathrm{pri}} := \mathbb{E}_{X|S=1}[D_{\mathrm{KL}}(p_{\pi_\theta}(Z \mid X, S=1) \| q_\tau(Z \mid X))]\,,$$

and

$$\epsilon^+_{\mathrm{post}} := \mathbb{E}_{X,R|S=1}[D_{\mathrm{KL}}(p_{\pi_\theta}(Z \mid X, R, S=1) \| q_\tau(Z \mid X, R))]\,.$$

Proof. Starting from equation 2, add and subtract $\log p_{\pi_\theta}(Z \mid X, R, S=1)$ and $\log p_{\pi_\theta}(Z \mid X, S=1)$ inside the expectation. This gives

$$\begin{aligned} I^+_\tau =& \mathbb{E}\left[\log \frac{p_{\pi_\theta}(Z \mid X, R, S=1)}{p_{\pi_\theta}(Z \mid X, S=1)}\right] \\ &+ \mathbb{E}\left[\log \frac{q_\tau(Z \mid X, R)}{p_{\pi_\theta}(Z \mid X, R, S=1)}\right] \\ &+ \mathbb{E}\left[\log \frac{p_{\pi_\theta}(Z \mid X, S=1)}{q_\tau(Z \mid X)}\right], \end{aligned}$$

where all expectations are under the corresponding distribution conditioned on $S=1$. The first term is I^+_{exe}, the second is $-\epsilon^+_{\mathrm{post}}$, and the third is $\epsilon^+_{\mathrm{pri}}$. Hence

$$I^+_\tau = I^+_{\mathrm{exe}} + \epsilon^+_{\mathrm{pri}} - \epsilon^+_{\mathrm{post}}.$$

Finally,

$$|I^+_\tau - I^+_{\mathrm{exe}}| = |\epsilon^+_{\mathrm{pri}} - \epsilon^+_{\mathrm{post}}| \le \epsilon^+_{\mathrm{pri}} + \epsilon^+_{\mathrm{post}},$$

which proves Proposition 1. □

B.7 Implications for the Practical Execution Signal

Proposition 1 does not require the frozen observer q_τ to exactly recover the actor-induced semantic distributions. Instead, the discrepancy between I^+_τ and I^+_{exe} is explicitly characterized by the observer's prior and posterior approximation errors. In particular,

$$I^+_\tau - (\epsilon^+_{\mathrm{pri}} + \epsilon^+_{\mathrm{post}}) \le I^+_{\mathrm{exe}} \le I^+_\tau + (\epsilon^+_{\mathrm{pri}} + \epsilon^+_{\mathrm{post}}).$$

The practical implementation of GroundESC operates on the textual Strategy Line T rather than directly on the analytical semantic abstraction Z. Accordingly, the token-level prior-to-posterior score used during training should be interpreted as a computable text-level surrogate motivated by I^+_τ, rather than as an unbiased estimator of I^+_{exe}. No equality between the practical text-level score and the semantic mutual-information quantity is required by the analysis above.

C Additional Experimental Details

C.1 Training Details

We instantiate GROUNDESC on four backbones from two model families: Gemma-2B, Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, and Qwen2.5-7B-Instruct. Unless otherwise specified, all backbones use the same training configuration. Both structured supervised fine-tuning and reinforcement learning update all model parameters without parameter-efficient adapters. All experiments are conducted in BF16 on a server equipped with eight NVIDIA L40S GPUs.

Structured supervised fine-tuning. We first perform structured supervised fine-tuning for one epoch. The model is optimized with AdamW using an initial learning rate of 3×10^{-5}, a cosine learning-rate schedule, a warmup ratio of 0.03, weight decay of 0.01, and gradient clipping at 1.0. The effective batch size is 16 sequences, and gradient checkpointing is enabled throughout training. The maximum input length is $4{,}096$ tokens.

The resulting SFT checkpoint initializes the three models used in the subsequent reinforcement-learning stage: the trainable actor π_θ, the frozen reference policy π_{ref}, and the frozen strategy execution scoring model π_τ. Thus, the actor, reference policy, and execution scorer share the same structured initialization before GRPO training.

GRPO training. Starting from the SFT checkpoint, we conduct full-parameter GRPO training for two epochs. Each training batch contains 32 dialogue histories, and the rollout policy samples $G = 8$ structured trajectories for each history, yielding 256 trajectories per batch. The actor learning rate is 5×10^{-7}, with one policy epoch per rollout batch. We use a clipping coefficient of 0.2 and a KL coefficient of 0.05 to regularize the actor toward the frozen reference policy.

The maximum context length during GRPO is $2{,}048$ tokens, and the maximum rollout length is 384 tokens. Rollouts are sampled with temperature 0.7, top-$p = 0.95$, and top-$k = 20$. During policy optimization, the strategy execution scoring model π_τ remains frozen and evaluates generated Strategy Lines using teacher-forced conditional likelihoods. The response-quality reward model is based on Qwen3-8B and is likewise kept fixed throughout reinforcement learning.

Reward configuration. The trajectory reward combines format compliance, contextual strategy selection, strategy execution, and response quality. Their weights are set to

$$(\lambda_f, \lambda_s, \lambda_e, \lambda_q) = (0.05,\ 0.25,\ 0.30,\ 0.40),$$

respectively. The KL penalty is applied separately from these trajectory-level reward components. Table 7 summarizes the principal training hyperparameters.

C.2 Data Construction

Structured supervised fine-tuning data. We construct the structured training set from SSConv and retain $19{,}849$ quality-controlled supporter turns. Each target follows the ordered form (C, T, R), where C analyzes the seeker's situation, affective state, and support needs; T is a textual Strategy Line specifying the focal support strategy and its intended supportive function; and R is the corresponding natural-language response. The original supporter response is retained as R, while the seeker-state analysis and Strategy Line provide supervision for the intermediate support decision. Following Section 2, the Strategy Line T represents the corresponding strategy intent $Z = \phi_X(T)$ used in our theoretical analysis.

Reinforcement learning data. We first verify sample identities, dialogue histories, structured outputs, and strategy labels in the supervised training pool. DeepSeek-V4-Flash (Xu et al., 2026) then assesses the candidate strategy categories for each dialogue history h. The resulting annotations define a set $\mathcal{A}^+(h) \subseteq \mathcal{A}$ containing between one and three contextually admissible strategy categories, together with a soft suitability score $\omega_h(a)$ for each admissible category. These annotations are stored only as reward-side metadata and are never included in the actor input.

Table 7: Training hyperparameters for structured supervised fine-tuning and GRPO.

Hyperparameter	**Structured SFT**	**GRPO**
Training epochs	1	2
Learning rate	3×10^{-5}	5×10^{-7}
Effective batch size	16 sequences	32 histories
Trajectories per history	–	8
Trajectories per batch	–	256
Policy epochs per batch	–	1
Warmup ratio	0.03	–
Weight decay	0.01	–
Gradient clipping	1.0	–
Clipping coefficient	–	0.2
KL coefficient	–	0.05
Maximum input/context length	4,096 tokens	2,048 tokens
Maximum rollout length	–	384 tokens
Rollout temperature	–	0.7
Top-p / top-k	–	0.95 / 20
Precision	BF16	BF16
Random seed	42	42

Because admissibility is annotated from the dialogue history before trajectory generation, for a generated context $x = (h, c)$ we use $\mathcal{A}^+(x) := \mathcal{A}^+(h)$ and $\omega_x(a) := \omega_h(a)$. This yields the admissibility set and suitability scores used by the strategy-selection reward in Section 3.2.

We further select contexts according to dialogue coverage, conversational position, strategy coverage, and contextual quality. The resulting reinforcement-learning set contains 7,883 contexts from 2,527 dialogues, including 2,613, 2,604, and 2,666 examples from the early, middle, and late portions of their respective dialogues. This provides balanced coverage across different stages of emotional-support conversations.

Quality reward model data. We construct response-preference pairs from SSConv contexts to train the quality reward model. DeepSeek-V4-Flash (Xu et al., 2026) is used to construct candidate pairs covering contextual relevance, empathic attunement, support effectiveness, and naturalness. Preferred responses provide contextually appropriate and effective emotional support, whereas paired negatives introduce controlled deficiencies while remaining fluent. We additionally include hard negatives exhibiting verbose, templated, repetitive, or mechanically supportive responses.

The reward-model training set contains 6,000 preference pairs, comprising 4,000 general-quality pairs and 2,000 hard-negative pairs. Separate development and audit sets contain 700 pairs each. All splits are constructed at the dialogue level to prevent contexts from the same conversation from appearing in different subsets.

C.3 Evaluation Protocol

Test set and generation. The evaluation set contains $N = 2{,}714$ target supporter turns from SSConv. Each example i is associated with an admissible strategy set $\mathcal{A}_i^+$ containing between one and three strategy categories. These annotations are used only for post-generation evaluation and are not included in the model input.

All models receive the same structured generation prompt and use greedy decoding with temperature 0, top-$p = 1$, top-$k = 0$, a maximum generation length of 768 tokens, and random seed 42. DeepSeek-V4-Flash (Xu et al., 2026) evaluates strategy execution and response quality through two separate evaluation channels. This evaluator is independent of the Qwen3-8B quality reward model used during training. All models use the same generation and evaluation configurations.

Strategy selection. For example i, let t_i denote the generated Strategy Line, $z_i = \phi_{x_i}(t_i)$ its declared strategy intent, and $\hat{a}_i = a(z_i)$ its focal strategy category. Strategy admissibility is computed

as

$$\mathrm{Adm} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{a}_i \in \mathcal{A}_i^+\}.$$

This metric evaluates the declared category against the complete admissible set rather than requiring an exact match with a single reference strategy.

Strategy execution. Strategy execution fidelity evaluates how clearly the final response expresses the supportive function declared by the generated Strategy Line. The execution evaluator observes the dialogue history, generated seeker-state analysis, textual Strategy Line, and final response, and assigns an execution score $\mathrm{SF}_i \in \{1, \ldots, 5\}$. A low score indicates that the declared strategy is weakly or not expressed in the final response, whereas a high score indicates clear and faithful execution. A strategy mentioned only in the analysis or Strategy Line is not considered executed unless the corresponding supportive behavior appears in the final response.

To jointly account for contextual strategy selection and execution, we report the Grounded Strategy Fidelity (GSF):

$$\mathrm{GSF} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{a}_i \in \mathcal{A}_i^+\}\,\mathrm{SF}_i.$$

Thus, an inadmissible, invalid, or missing strategy declaration contributes zero to GSF, while an admissible strategy is weighted by how faithfully its declared intent is executed in the response.

Response quality. Response quality is evaluated independently along four dimensions: **contextual relevance** (CR), whether the response addresses the seeker's current situation and support needs; **empathic attunement** (EA), whether it appropriately recognizes and responds to the seeker's emotional state; **support effectiveness** (SE), whether it provides suitable and helpful support; and **naturalness** (Nat.), whether it is fluent, specific, and free from rigid or repetitive expressions.

The quality evaluator observes only the dialogue history and final response. It does not receive the generated seeker-state analysis, Strategy Line, or strategy-evaluation annotations, preventing the declared strategy from directly influencing response-quality judgments.

Each dimension is scored on a five-point scale. We report the four dimensions individually and summarize them as

$$\mathrm{Q4} = \frac{1}{4}\left(\mathrm{CR} + \mathrm{EA} + \mathrm{SE} + \mathrm{Nat}\right). \tag{1}$$

Because strategy frequencies are imbalanced, we additionally report Macro Q4, obtained by first computing Q4 within each focal strategy category and then averaging equally across categories.

Format validity. Format validity is evaluated mechanically. A generation is considered valid when the seeker-state analysis, Strategy Line, and final response are all present, nonempty, correctly ordered, and the Strategy Line contains exactly one parsable focal strategy declaration.

C.4 Calibration of the Strategy Execution Reward

As discussed in Section 3.2, the scale of the raw information score m_i may vary systematically across strategy categories. Some strategies express their supportive functions more explicitly in language or are more readily recognized by the frozen scorer, resulting in different score ranges even under comparable execution quality. We therefore calibrate m_i separately for each focal strategy category.

We estimate the calibration statistics offline using samples constructed exclusively from the training set. Let $\mathcal{D}_{\mathrm{cal}}$ denote the calibration set. For each sample (x_j, t_j, r_j), let $z_j = \phi_{x_j}(t_j)$ denote the strategy intent represented by its Strategy Line. For each category $k \in \mathcal{A}$, we collect

$$\mathcal{M}_k = \{m_j \mid (x_j, t_j, r_j) \in \mathcal{D}_{\mathrm{cal}},\ a(z_j) = k\}. \tag{2}$$

We define two category-specific reference points as

$$b_k = Q_{\rho_b}(\mathcal{M}_k), \qquad u_k = Q_{\rho_u}(\mathcal{M}_k), \qquad \rho_b < \rho_u, \tag{3}$$

Table 8: Human evaluation of the strategy execution signal on Qwen2.5-7B.

Dimension	ρ_s (95% CI)	κ_w	α_H	MAE
CR	0.850 [0.807, 0.885]	0.777	0.675	0.118
EA	0.905 [0.884, 0.922]	0.779	0.683	0.142
SE	0.911 [0.891, 0.929]	0.798	0.730	0.143
Nat.	0.871 [0.842, 0.894]	0.753	0.665	0.146
SF	0.764 [0.713, 0.807]	0.654	0.603	0.257

where $Q_\rho(\cdot)$ denotes the empirical ρ-quantile. The lower reference b_k represents the category-specific baseline, while u_k represents a high level of response evidence. The quantile levels are reported in the implementation details.

The category-specific calibration function is

$$g_k(m) = \text{clip}\left(\frac{m - b_k}{u_k - b_k}, 0, 1\right). \tag{4}$$

Accordingly, the strategy execution reward for trajectory i is

$$e_i = g_{a(z_i)}(m_i). \tag{5}$$

This calibration evaluates the information score relative to the distribution of its focal strategy category, avoiding a shared raw-score scale across strategies. The statistics $\{b_k, u_k\}_{k\in\mathcal{A}}$ are estimated before reinforcement learning and remain fixed throughout policy optimization. Contextual admissibility is incorporated separately when the execution reward is combined with the other rewards.

D More Results

D.1 Human Validation of LLM-Based Evaluation

Several components of our experimental protocol rely on LLM-based judgments. We therefore conduct two human validation studies to assess their agreement with human annotations. The first examines the evaluator used for strategy execution and response-quality assessment, while the second evaluates LLM-generated strategy-admissibility annotations used for strategy selection.

Agreement on response evaluation. We first compare human and LLM scores on 400 responses sampled from 200 dialogue clusters. The evaluation covers contextual relevance (CR), empathic attunement (EA), support effectiveness (SE), naturalness (Nat.), and strategy fidelity (SF). We report Spearman correlation ρ_s, weighted Cohen's κ_w, human annotation reliability α_H, and mean absolute error (MAE).

As shown in Table 8, the LLM evaluator aligns closely with human judgments across all five dimensions, providing empirical support for its use in large-scale evaluation. For the four response-quality dimensions, Spearman correlations range from 0.850 to 0.911, with weighted κ between 0.753 and 0.798. Strategy fidelity also exhibits substantial human–LLM agreement, with $\rho_s = 0.764$ and $\kappa_w = 0.654$, supporting the reliability of the evaluator for assessing strategy execution.

Agreement on strategy admissibility. We separately validate the strategy-admissibility annotations used by the selection objective and evaluation protocol. For 200 dialogue contexts, human annotators independently identify all strategy categories that are contextually admissible. We compare these labels with the corresponding LLM annotations at the category level and report the positive-label frequency, agreement coefficient, and F1 score.

Table 9 demonstrates strong agreement between human and LLM strategy-admissibility annotations, with pooled $\kappa_H = 0.876$ and F1 $= 0.899$. The LLM annotations closely reproduce human admissibility judgments across the major strategy categories, including Question, Reflection, Affirmation, and Suggestions. This agreement provides direct human validation of the strategy-selection supervision used in GroundESC, supporting the reliability of both the selection objective and the corresponding large-scale evaluation protocol.

Table 9: Agreement between human and LLM strategy-admissibility annotations on 200 dialogue contexts. "Human +" and "LLM +" denote the proportion of contexts in which each strategy is labeled admissible.

Strategy	**Human +**	**LLM +**	κ_H	**F1**
Question	18.5%	17.5%	0.932	0.944
Restat./Para.	1%	0.05%	0.664	0.667
Reflection	26.0%	25.5%	0.778	0.835
Self-discl.	1.5%	2.0%	0.855	0.857
Affirmation	48.0%	47.0%	0.820	0.905
Suggestions	32.0%	31.0%	0.907	0.937
Information	2.5%	4.0%	0.762	0.769
Pooled	–	–	**0.876**	**0.899**

Table 10: Fully human evaluation of the strategy execution signal on Qwen2.5-7B. All variants are evaluated on the same 200 dialogue contexts. F, Q, S, E denote format, quality, selection, and execution rewards.

Backbone	**Training**	**Reward**				**Support-Strategy Grounding**			**Response Quality**			
		F	Q	S	E	**Adm. (%)↑**	**SF↑**	**GSF↑**	**CR↑**	**EA↑**	**SE↑**	**Nat.↑**
Qwen2.5-7B-Instruct	SFT	–	–	–	–	31.50	3.327	1.049	3.602	3.212	3.178	3.367
	RL	✓	✓	–	–	37.50	3.487	1.311	3.825	3.496	3.382	3.416
	RL	✓	✓	✓	–	59.00	3.368	1.987	3.855	3.398	3.410	3.491
	RL	✓	✓	✓	✓	**60.50**	**4.022**	**2.457**	**4.068**	**3.658**	**3.530**	**3.723**

D.2 Fully Human Validation of Grounding Objectives

To verify that the gains of GroundESC are not specific to the automatic annotation and evaluation pipeline, we conduct a fully human ablation study on the same 200 dialogue contexts. Human annotators independently identify the contextually admissible strategy set for each dialogue and evaluate strategy fidelity (SF), contextual relevance (CR), empathic attunement (EA), support effectiveness (SE), and naturalness (Nat.). Accordingly, all grounding and response-quality metrics in Table 10, including Adm. and GSF, are derived entirely from human judgments.

Table 10 reproduces the intended effects of the grounding objectives under fully human evaluation. Adding the strategy-selection reward S increases admissibility from 37.50% to 59.00%, while strategy fidelity remains largely unchanged, indicating that S primarily improves contextually appropriate strategy choice. In contrast, adding the execution reward E leaves admissibility nearly unchanged ($59.00\% \rightarrow 60.50\%$) but substantially improves SF ($3.368 \rightarrow 4.022$) and GSF ($1.987 \rightarrow 2.457$). This complementary pattern directly matches the intended roles of the two objectives: S strengthens context-to-strategy grounding, whereas E strengthens strategy-to-response realization.

The full model also achieves the best human-rated scores on all four response-quality dimensions. Thus, improving strategy grounding does not require sacrificing overall response quality. More importantly, the same selection–execution decomposition observed in the automatic evaluation persists under independent human judgments, providing direct evidence that the gains of GroundESC reflect the targeted grounding behavior rather than artifacts of the automatic supervision or evaluation procedure.