

From Expressions to Needs: Hierarchical Interpretive Reasoning for Emotional Support Conversation

Xinjie Zhang^{1,*}, Haotian Fan^{1,*}, Zhaopei Huang¹,
Shujie Ren², Wenxuan Wang¹, Qin Jin^{1,†}

¹School of Information, Renmin University of China, ²Beijing Forestry University
{zhangxinjie827, wangwenxuan, qjin}@ruc.edu.cn

Abstract

Existing emotional support conversation (ESC) systems often react directly to surface-level distress expressions, generating reassurance or advice without adequately modeling the underlying psychological needs driving the distress. However, effective emotional support requires an interpretive process that connects *what is expressed* with *why it is expressed* and ultimately *what support is needed*. We argue that current ESC systems lack such an *interpretive layer* between observation and supportive intervention. To address this gap, we propose **HierINT**, a psychologically grounded hierarchical interpretive schema for emotional support conversations. HierINT distinguishes observable communicative intentions from the underlying psychological interpretations that drive them, including cognitive appraisals and unmet fundamental psychological needs. Building on this schema, we construct **HierINT-Dlg**, a dataset with hierarchical interpretive annotations and need-aligned support strategies. We further propose **HierINT-CoRe**, a collaborative reasoning framework that operationalizes HierINT by jointly modeling psychological interpretation and support strategy planning prior to response generation. Experiments on ESC-Conv and HierINT-Dlg demonstrate consistent improvements over surface-level and intention-aware baselines, particularly on empowerment and personalization metrics. Our results suggest that effective emotional support requires moving beyond reactive response generation toward psychologically guided, need-informed intervention. Our data and code are available at <https://github.com/43zxj/HierINT>.

1 Introduction

Emotional Support Conversation (ESC) aims to provide empathetic and psychologically meaningful support to individuals experiencing emotional

distress (Albrecht and Adelman, 1987; Burleson, 2003). Unlike task-oriented dialogue systems (Hu et al., 2023; Zhang et al., 2024a) or open-domain conversation agents (Li et al., 2017), ESC requires understanding not only what users say, but also the underlying psychological states and unmet needs driving those expressions. Recent large language models (LLMs) have substantially improved the fluency and empathy of supportive responses (Zhang et al., 2024b, 2025). However, most existing ESC systems still operate within a reactive *recognize-and-respond* paradigm: once distress is detected, models immediately generate reassurance, validation, or advice based primarily on surface-level expressions.

We argue that this paradigm overlooks a critical component of effective emotional support: **interpretation before intervention**. In human supportive interaction, effective support rarely emerges from reacting directly to observable expressions alone. Instead, supporters typically infer why a person feels distressed, what psychological meanings they assign to the situation, and what unmet needs underlie their emotion before deciding how to help. For example, two seekers may both express frustration and self-doubt, yet one may primarily need validation of competence, while another may require restoration of autonomy or emotional reassurance. Without modeling these underlying psychological differences, systems may produce superficially empathetic but psychologically misaligned responses.

Figure 1 illustrates this limitation. A seeker states: *"I've been working so hard, but my boss still criticized my report. I feel like I'm just not good enough."* Existing ESC systems often interpret this expression primarily as reassurance-seeking and respond with generic encouragement such as *"You're doing great!"* However, the same surface expression may reflect substantially different underlying psychological structures. From the perspective of Cognitive Appraisal Theory (Lazarus, 1984), the

*Co-first authors with equal contribution.

†Corresponding Author.

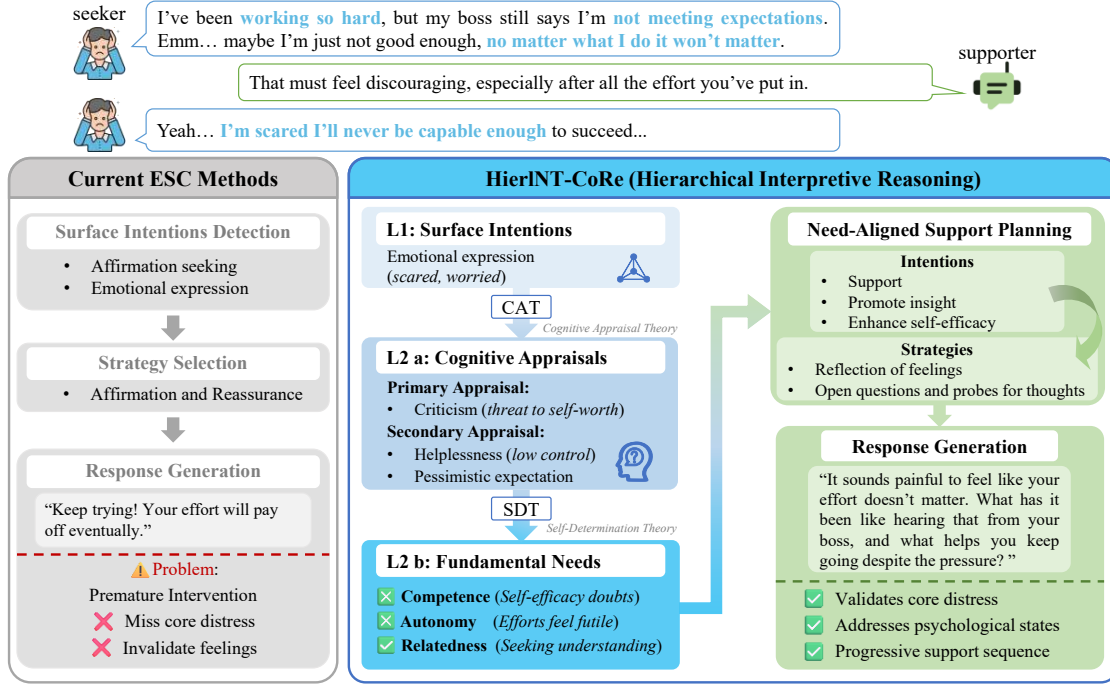


Figure 1: Comparison between existing ESC systems and our HierINT-CoRe framework. Existing approaches often directly map observable distress expressions to supportive responses, while HierINT-CoRe introduces an interpretive layer that connects surface expressions with underlying psychological interpretation and need-aligned support planning.

seeker may interpret criticism as a threat to self-worth and perceive limited control over the situation. Self-Determination Theory (Ryan and Deci, 2000) further suggests unmet needs related to competence, autonomy, or relatedness. Consequently, effective support requires more than immediate reassurance: it requires interpreting the seeker’s underlying psychological needs before selecting supportive strategies. This example highlights a fundamental challenge in ESC: emotional support should not be modeled as a direct reaction, but as an interpretive process connecting observable expressions with psychologically appropriate intervention.

Unfortunately, current ESC systems largely lack such an interpretive layer. At the **representation level**, existing approaches typically model seeker intentions as flat communicative labels derived from isolated utterances (Cai and Chen, 2020; Kotnis et al., 2022; Kuo and Chen, 2023). Although these methods capture observable conversational goals, they often conflate what seekers explicitly express with what they psychologically need. At the **reasoning level**, recent intention-aware ESC systems (Zhang et al., 2025) and chain-of-thought reasoning approaches (Zhang et al., 2024b) improve supportive planning, yet still focus primarily on deciding *what* action to take rather than reason-

ing about *why* a particular intervention is psychologically appropriate. As a result, existing systems remain largely reactive, generating support from surface patterns rather than grounded psychological interpretation.

To address these limitations, we propose **HierINT**, a psychologically grounded hierarchical interpretive schema for emotional support conversations. Rather than directly mapping observable expressions to supportive responses, HierINT introduces an intermediate interpretive layer that connects *what is expressed* with *why it is expressed* and ultimately *what support is needed*.

Specifically, HierINT distinguishes two interconnected layers: **L1: Surface Intentions**, capturing observable communicative goals such as emotional disclosure, reassurance seeking, or advice seeking; **L2: Underlying Psychological Interpretation**, modeling the latent psychological structures underlying these expressions, including cognitive appraisals, emotional states, and unmet fundamental psychological needs. Importantly, HierINT reframes emotional support as a need-informed interpretive process rather than reactive response generation. Supportive actions are therefore determined not merely by what seekers explicitly ask for, but by the underlying psychological needs inferred beneath surface-level expressions.

Building upon this schema, we construct **HierINT-Dlg**, a dataset that includes turn-level annotations for surface intentions, psychological interpretations, supporter intentions, and psychologically grounded support strategies. We further propose **HierINT-CoRe**, a collaborative reasoning framework that operationalizes HierINT by jointly modeling psychological interpretation and support strategy planning prior to response generation.

We position HierINT as both a *representational* and *reasoning-level* contribution. We argue that current ESC systems underperform not merely because of insufficient generation capability, but because they lack the appropriate reasoning structure between observation and intervention. HierINT provides this missing interpretive structure, while HierINT-CoRe demonstrates that operationalizing psychologically grounded interpretation yields consistent improvements across multiple backbones, evaluators, and datasets.

Our main contributions are summarized as follows: (1) We propose HierINT, a hierarchical interpretive schema that introduces an explicit interpretive layer between surface observation and supportive intervention in ESC. (2) We construct a dataset with hierarchical interpretive annotations and need-aligned support strategies for emotional support conversations. (3) We propose HierINT-CoRe, a collaborative reasoning framework for joint psychological interpretation and strategy planning, and empirically demonstrate its consistent superiority over reactive generation approaches across multiple backbones and evaluation dimensions.

2 HierINT Schema

Having identified the missing interpretive layer as the core limitation of existing ESC systems, we now formalize what this layer should contain. Skilled counselors do not respond to surface expressions directly; instead, they reason through multiple layers—from what is expressed, to why it is expressed, to what kind of support is psychologically needed (Hill and O’Grady, 1985). HierINT formalizes this process as a psychologically grounded hierarchical interpretive schema for emotional support conversations.

Rather than treating emotional support as direct reaction to observable distress, HierINT introduces an intermediate interpretive layer between observation and intervention. The schema decomposes seeker understanding into two intercon-

nected levels: (i) **Surface Intentions**, capturing the seeker’s observable communicative goals and support-seeking behaviors; and (ii) **Underlying Psychological Interpretation**, modeling the latent psychological structures that explain these expressions, including cognitive appraisals, emotional states, and unmet psychological needs. Within this hierarchy, surface intentions characterize *what* the seeker expresses, while underlying psychological interpretation explains *why* the expression arises and *what support is fundamentally needed*.

2.1 Surface Intentions

Surface intentions capture the seeker’s observable communicative behaviors and conversational goals. Following Hill’s Client Behavior System (CBS) (Hill et al., 1992), we define seven surface intention categories (e.g., *Emotional Expression*, *Recounting*) that characterize the functional role of an utterance in the conversation.

However, surface intentions alone remain insufficient for effective emotional support because they describe only the form of expression (*what* is communicated), rather than the underlying psychological drivers (*why* it is communicated). For example, a seeker recounting a workplace conflict may be seeking emotional validation, requesting practical guidance, processing self-doubt, or attempting to regain a sense of control. Although these cases may appear similar at the conversational surface level, they imply substantially different psychological needs and therefore require different supportive interventions. This limitation motivates the need for a deeper interpretive layer.

2.2 Underlying Psychological Interpretation

Underlying psychological interpretation models the latent psychological structures driving surface expressions. A principled interpretive framework requires theoretical grounding; otherwise, deeper reasoning risks degenerating into ad hoc speculation.

We use two complementary psychological theories to define the operational categories and ordered supervision targets of this layer, rather than architectural constraints. First, Lazarus’s Cognitive Appraisal Theory (Lazarus, 1984) explains how emotions emerge through subjective evaluations of situations (Lazarus and Smith, 1988). We model five appraisal dimensions: *Self-relevance* and *Self-congruence* (primary appraisal), together with *Blame or Credit*, *Coping Potential*, and *Future Expectations* (secondary appraisal). These dimen-

sions characterize how seekers cognitively interpret stressful experiences and evaluate their ability to respond. While cognitive appraisals explain why distress arises, they do not directly specify what type of support is psychologically needed. For instance, a seeker with low perceived coping potential may require emotional validation, autonomy-respecting guidance, or competence-oriented encouragement depending on the underlying unmet need. To model this need-oriented dimension, we further incorporate Self-Determination Theory (SDT) (Ryan and Deci, 2000), which identifies three fundamental psychological needs: *Autonomy*, *Relatedness*, and *Competence*.

We adopt SDT as a principled balance between theoretical expressiveness and operational tractability. Compared with broader motivational taxonomies, SDT provides: (i) empirically validated dimensions associated with psychological well-being across diverse populations (Ryan and Deci, 2000); (ii) sufficiently distinct categories to support reliable annotation; and (iii) direct relevance to supportive intervention, since different unmet needs correspond to different support strategies. Finer-grained extensions remain a promising future direction once annotation reliability at higher semantic granularity becomes feasible.

By integrating cognitive appraisal with need-oriented interpretation, this layer bridges observable expressions and psychologically grounded support understanding. Detailed annotation guidelines are provided in Appendix B.2.

2.3 From Psychological Interpretation to Support Actions

Once underlying psychological interpretations are inferred, the supporter’s task shifts from understanding to intervention planning. Importantly, the same surface expression may require fundamentally different support depending on the inferred psychological need. For example, frustration associated with threatened competence may benefit from efficacy-affirming encouragement, whereas frustration rooted in unmet relatedness may require emotional validation and relational reassurance.

Drawing from counseling practices (?) and prior ESC research (Zhang et al., 2025), we define 14 supporter intentions (e.g., *Promote Insight*, *Get Information*), each associated with different psychological support functions. Because supportive behavior is often multi-intentional, each supporter intention may be realized through different need-

aligned strategies (Zhang et al., 2025). This completes the interpretive reasoning pathway formalized by HierINT: surface expressions → cognitive appraisals → unmet psychological needs → need-aligned supporter intentions → supportive strategies. Detailed definitions of supporter intentions and support strategies are provided in Appendix B.3 and Appendix B.4.

3 HierINT-Dlg Dataset

HierINT defines the representational target; HierINT-Dlg provides the empirical grounding. Existing ESC datasets (Liu et al., 2021; Zheng et al., 2023b; Zhang et al., 2024b; Chen et al., 2025; Zhang et al., 2025) do not distinguish between partially and fully articulated seeker intentions, nor provide supervision for reasoning about deeper needs. We construct **HierINT-Dlg**, a dialogue dataset with hierarchical interpretive annotations, designed to support research on uncovering latent needs and understanding the rationale behind effective supporter responses.

3.1 Dialogue Generation

Inspired by recent simulation-based efforts (Chen et al., 2025), we build a large-scale seeker persona pool combining demographic attributes and Big Five personality profiles (Goldberg, 1992), with situational contexts sampled from Zhang et al. (2024b) and generated by GPT-4.1 (OpenAI, 2025). Seeker utterances are conditioned on persona, situation, and dialogue history, producing diverse linguistic and emotional expressions (details in Appendix C.1). On the supporter side, following Hill and O’Grady (1985); Zhang et al. (2025), utterances are produced via a structured three-step framework: (1) determine intention(s), (2) select corresponding strategy(ies), and (3) generate a coherent response. Multiple complementary intentions are encouraged per utterance, mirroring how skilled counselors simultaneously pursue multiple support goals. Human pairwise evaluation confirms our approach outperforms the strategy-only baseline (31 wins vs. 15 losses, 4 ties; see Appendix C.3).

3.2 Hierarchical Interpretive Reasoning Annotation

We annotate both sides of the reasoning process—hierarchical seeker understanding that reveals psychological drivers, and the complete supporter reasoning chain that translates this understanding

into targeted actions—on sampled dialogue snippets of varying lengths and interaction stages (Appendix C.6).

Seeker intention annotation. Surface intentions are annotated from the most recent utterance. Underlying psychological interpretation follows a four-stage reasoning process: *Event Identification (EI) → Primary Appraisal (PA) → Secondary Appraisal (SA) → Underlying Needs (UN)*, forming an interpretive chain from surface cues to psychological drivers. **Supporter reasoning annotation.** Building on seeker chains and existing intention-strategy labels, we add an *Intention Explanation* linking supporter choices to seeker needs (prompt in Figure 8), providing supervision for the full reasoning flow. A complete case is in Figure 14.

3.3 Quality Control

To ensure data safety and annotation reliability, we implement comprehensive quality control throughout all stages. For **realism and diversity**, demographics follow real-world distributions (United Nations, 2024) with ISCO-08 occupations (International Labour Office, 2012), and a two-stage safety mechanism ensures persona-topic compatibility while prohibiting extreme scenarios. Models trained on HierINT-Dlg generalize effectively to the independently collected ESConv benchmark (Table 2), whose dialogues were produced by real crowd-workers, suggesting transferable conversational patterns rather than dataset-specific artifacts. For **annotation quality**, trained annotators review dialogue snippets with intention and strategy labels, refining 20% of intention annotations and 14% of strategy annotations (Appendix C.4), confirming that synthetic annotations align with human expert judgment.

Category	HierINT-Dlg	ESConv
# Dialogues	2,674	1,300
# Utterances	100,944	38,365
Avg Dialogue Len (# Uttr)	37.8	29.5
Avg Uttr Len (# Words)	30.3	16.4
Avg Seeker Uttr Len	32.7	14.8
Avg Supporter Uttr Len	28.8	18.1
Supporter Avg Intentions	1.88	-
Supporter Avg Strategies	1.44	1.00

Table 1: Statistical comparison of dialogues in HierINT-Dlg and ESConv.

3.4 Statistics

HierINT-Dlg consists of 2,674 full dialogues with multi-intention/strategy labels and 752 dialogue snippets with enriched hierarchical interpretive annotations (Table 1). Each snippet includes a complete underlying psychological interpretation chain ($EI \rightarrow PA \rightarrow SA \rightarrow UN$) and full supporter reasoning chains, supporting training for both intention mining and intention-guided generation. Further analysis in Appendix C.6.

4 HierINT-CoRe Framework

HierINT defines *what* to model; HierINT-Dlg provides *what to learn from*; HierINT-CoRe defines *how* to operationalize the inference at scale. We propose **HierINT-CoRe**, a **Collaborative Reasoning** framework that operationalizes the HierINT schema: inferring from observable expressions to underlying psychological states, then translating identified needs into targeted support. As illustrated in Figure 2, HierINT-CoRe consists of three coordinated modules: (1) **Seeker Psychological Interpretation**, which progressively uncovers underlying needs from surface expressions; (2) **Supporter Intention Reasoning**, which translates identified needs into support goals; and (3) **Strategy Selection and Response Generation**, which realizes these goals through coherent multi-strategy responses. The resulting improvements generalize across different backbone models (LLaMA3.1, Qwen3; see Appendix D.14).

4.1 Seeker Psychological Interpretation Module

This module infers the seeker’s latent psychological state via hierarchical reasoning. Let the dialogue context up to turn t be $D_t = \{(u_1^s, u_1^p), \dots, (u_{t-1}^s, u_{t-1}^p), u_t^s\}$, where u_i^s and u_i^p represent the seeker’s and supporter’s utterances at turn i , and u_t^s denotes the seeker’s current utterance. The **surface intention** $I_t^{\text{surf}} \subset \mathcal{I}_{\text{surf}}$ refers to a subset of predefined observable communicative goals or support-seeking behaviors expressed in the seeker’s current utterance u_t^s . The **underlying psychological interpretation** $I_t^{\text{upi}} = \{\Psi_t, UN_t\}$ encompasses two components: the cognitive appraisal state $\Psi_t = \{EI, PA, SA\}$, and the unmet fundamental needs UN_t inferred from the dialogue context D_t . We formalize the module’s task as a sequence-to-structure mapping problem, where the input dialogue history is mapped to a structured

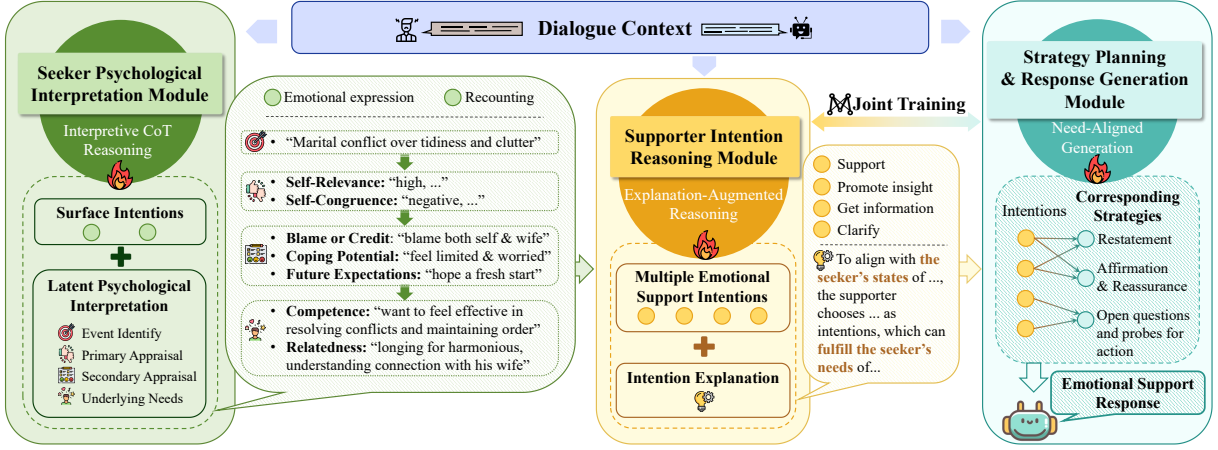


Figure 2: Framework of the proposed HierINT-CoRe.

representation of the seeker’s intentions:

$$I_t^{\text{surf}}, I_t^{\text{upi}} = f_{\text{seeker}}(D_t) \quad (1)$$

To operationalize the surface-to-depth reasoning process, we employ Chain-of-Thought (Wei et al., 2022) reasoning. Specifically, the cognitive appraisal state Ψ_t acts as an intermediate reasoning step: it guides the inference of the seeker’s underlying emotional needs UN_t , thereby improving both the interpretability and precision of psychological interpretation.

4.2 Supporter Intention Reasoning Module

This module translates mined seeker intentions into specific support goals. Since effective support is inherently **multi-intentional**—skilled supporters simultaneously pursue complementary goals—the module predicts the set of supporter intentions $I_t^p \subset \mathcal{I}_p$ and generates a natural language Explanation IE_t articulating the reasoning behind them (Narang et al., 2020):

$$I_t^p, IE_t = f_{\text{supporter}}(D_t, I_t^{\text{surf}}, I_t^{\text{upi}}) \quad (2)$$

4.3 Supporter Strategy Planning and Response Generation Module

The final module operationalizes supporter intentions through concrete strategies and coherent responses. To handle **multi-intention coordination**, it follows a structured process: (1) group co-occurring intentions, (2) select strategies for each group, and (3) generate targeted responses. Formally:

$$\{(S_{t,i}, R_{t,i})\}_{i=1}^k = f_{\text{gen}}(D_t, I_t^{\text{surf}}, I_t^{\text{upi}}, I_t^p) \quad (3)$$

where $S_{t,i} \subset \mathcal{S}$ denotes a subset of strategies, and $R_{t,i}$ is the response designed to realize these strategies. The final response R_t is not a simple concatenation of $R_{t,i}$; instead, it integrates these units into a cohesive output that reflects the supporter’s multi-intention reasoning. This design ensures all intentions are addressed coherently, enabling precise, comprehensive, and emotionally adaptive support. In practice, we jointly train the Supporter Intention Reasoning and Strategy–Response Generation modules, as this joint optimization yields the best performance (see experiments).

5 Experiments

Our evaluation validates the central hypothesis from multiple angles: baseline comparison with controlled backbone, ablation of each hierarchy level, interactive multi-turn evaluation, human evaluation of intermediate outputs, and cross-evaluator/multi-backbone robustness checks.

5.1 Experimental Setting

Baselines and Datasets We compare HierINT-CoRe with several baselines: recent LLM-based approaches ESCoT (Zhang et al., 2024b) and IntentionESC (Zhang et al., 2025), both re-implemented using the same backbone (LLaMA3.1-8B-Instruct (Grattafiori et al., 2024)) as ours for fair comparison, plus BlenderBot-joint (Liu et al., 2021) and MultiESC (Cheng et al., 2022) as historical reference points. Experiments and evaluations are conducted on ESConv (Liu et al., 2021) (8:1:1 split) and HierINT-Dlg (7:1:2 split).

Evaluation Methods We evaluate response quality from two perspectives: basic quality (*Fluency*, *Coherence*, *Consistency*, *Safety*) following Zhang et al. (2025), and emotional support quality—extending

Datasets	Models	Basic Quality				Emotional Support Quality				
		Flue.	Cohe.	Cons.	Safe.	Empa.	Info.	Empo.	Pers.	Overall
ESConv	BlenderBot-Joint	4.08	2.98	3.32	2.69	2.03	1.20	1.33	0.67	1.31
	MultiESC	3.56	2.71	3.07	2.64	1.85	1.12	1.29	0.61	1.22
	ESCoT	4.84	4.22	4.48	3.49	3.33	1.59	2.19	1.22	2.08
	IntentionESC	4.83	3.82	4.18	3.12	2.41	1.35	1.57	0.82	1.54
	HierINT-CoRe	4.70	4.71	4.80	3.69	3.72	1.95	2.53	1.68	2.47
HierINT-Dlg	BlenderBot-Joint	2.03	1.98	2.36	2.41	2.41	1.19	1.65	0.89	1.54
	ESCoT	5.00	4.20	4.63	3.49	3.69	2.08	2.83	1.86	2.62
	IntentionESC	4.98	3.66	4.03	3.29	2.92	1.65	2.25	1.45	2.07
	HierINT-CoRe	4.98	4.73	4.92	4.00	4.15	2.64	3.53	2.72	3.26
	- w/o SI(Interp.)	4.97	4.73	4.93	3.78	4.07	2.58	3.52	2.53	3.18
	- w/o SI	4.99	4.75	4.91	3.87	4.09	2.63	3.43	2.53	3.17
	- w/o PI	4.98	4.64	4.81	3.83	4.08	2.49	3.44	2.67	3.17
	- w/o ALL	4.91	4.71	4.88	3.76	4.04	2.50	3.32	2.39	3.06

Table 2: Performance comparison of different models on ESConv and HierINT-Dlg datasets. The ablation study is conducted on HierINT-CoRe with the LLAMA3.1-8B-INSTRUCT backbone to investigate the impact of each module. SI: Seeker Intention, PI: Supporter Intention.

Empathy (Zhao et al., 2023) and *Informativeness* (Zhou et al., 2023) with two novel dimensions: *Empowerment*, measuring whether responses acknowledge efforts and provide structured action pathways; and *Personalization*, reflecting situation-specific tailoring (details in Appendix D.3). All eight metrics are rated by DeepSeek-V4-Flash on a 5-point Likert scale (Likert, 1932), with human evaluation verifying reliability (Appendix D.5).

5.2 Comparison with Baselines

Core question: Does the interpretive layer yield better support than systems operating without it?

Table 2 presents a performance comparison between HierINT-CoRe and baselines (cases in Appendix D.10). Our approach achieves gains across all emotional support dimensions, with the largest improvements on *personalization* and *empowerment*—dimensions sensitive to interpretive depth—suggesting that need-informed reasoning enables responses tailored to psychological profiles. To assess statistical reliability, HierINT-CoRe outperforms the strongest baseline on all 8 dimensions across both datasets ($p < 0.004$ by one-sided sign test), confirming that the improvement pattern is unlikely due to chance. *Informativeness* scores remain relatively low across all models, consistent with the nature of emotional support dialogues where single-turn responses prioritize comfort and reflection over factual information. To mitigate potential style-bias from GPT-4.1-assisted data construction, our primary evaluator (DeepSeek-V4-Flash) is independent from the data generation pipeline; cross-validation with GPT-4.1 confirms consistent rankings, and Qwen3-8B as an alter-

native backbone yields the same relative ordering (Appendix D.14). Human correlation analysis (Appendix D.5, Spearman’s $\rho = 0.33\text{--}0.49$, all $p < 0.01$) further confirms that HierINT-CoRe consistently ranks first across all evaluation settings.

5.3 Ablation Study

Core question: Which level of the interpretive hierarchy contributes most to support quality?

To isolate each component’s impact, we conduct an ablation study (Table 2). Removing any module reduces performance, confirming the necessity of all components. Specifically, omitting the psychological interpretation layer—the interpretive core of HierINT—significantly reduces *informativeness*, *empowerment*, and *personalization*, confirming that modeling underlying needs is the primary driver of improvement. Removing supporter intention modeling also leads to substantial drops, consistent with Zhang et al. (2025). While individual ablation differences are modest in absolute terms—typical of systems where components are designed to be *complementary* rather than independently sufficient—two observations collectively establish significance: (i) the direction of degradation is *perfectly consistent* across all emotional support dimensions for every ablation ($p < 0.004$ by sign test across 8 dimensions); and (ii) the largest drops concentrate on *personalization* and *empowerment*—precisely the dimensions most theoretically linked to need-level reasoning—providing construct validity beyond mere numerical gains. These results confirm that the *depth* of understanding (underlying needs), not merely the *breadth* of modeling, drives meaningful improvement. A representative

ablation case is in Appendix D.11.

5.4 Interactive Dialogue Evaluation

Core question: Does need-informed reasoning sustain its advantage over multi-turn interactions?

We employ ESC-Eval (Zhao et al., 2024), a role-play-based ESC evaluation framework. Following its standard protocol, we compare across dialogue lengths of 5, 8, 10, and 15 turns with 10 role cards per setting (40 dialogues total per model). We additionally include MultiAgentESC as it represents a recent multi-agent paradigm, which is excluded from the main comparison (Table 2) as its multi-agent architecture is not directly comparable under our controlled single-backbone setting. As shown in Table 3, HierINT-CoRe achieves the highest overall score with consistent improvements in *Expression*, *Information*, and *Skill*, while maintaining comparable *Fluency* and *Empathy*. The consistent advantage across various dialogue lengths suggests that the pattern is robust rather than driven by specific scenarios.

Models	Flu.	Exp.	Emp.	Inf.	Hum.	Ski.	Overall
ESCoT	2.85	2.78	3.0	2.72	2.25	2.90	1.85
IntentionESC	3.0	2.65	3.0	2.40	2.30	2.70	1.75
MultiAgentESC	3.0	2.83	3.0	2.75	2.45	2.85	1.85
HierINT-CoRe	3.0	2.95	3.0	2.9	2.33	2.95	1.98

Table 3: Performance comparison on ESC-Eval.

5.5 Seeker State Interpretation Evaluation

Core question: Does HierINT produce better psychological interpretations of seeker state?

We conduct human evaluation comparing HierINT with other state-inference approaches on a 3-point scale across five dimensions: *Groundedness*, *Psychological Depth*, *Explanatory Adequacy*, *Actionability*, and *Coverage* (criteria in Appendix D.4). As shown in Table 4, HierINT captures deeper psychological states, producing analyses that clarify *why* a seeker expresses an utterance and provide more actionable guidance. The performance gap is most pronounced on *Depth* and *Actionability*, where baselines score substantially lower, confirming that HierINT’s advantage lies specifically in structured psychological reasoning rather than surface-level completeness.

5.6 Joint Modeling of Supporter Tasks

Core question: Can the full supporter reasoning chain be unified without sacrificing quality?

Models	Grounded	Depth	Explain	Action	Coverage
ESCoT	2.92	2.33	2.33	2.30	2.48
IntentionESC	2.23	1.33	1.35	1.82	1.75
HierINT	2.95	2.97	2.98	2.93	2.92

Table 4: Human evaluation results on seeker state inference. The mean Krippendorff’s α across all dimensions is 0.734, and details are in Appendix D.4

Supporter-side tasks include intention reasoning, strategy selection, and response generation. We compare separate models for each task versus a single jointly-trained model. While separate models achieve competitive scores across emotional support quality dimensions (*Empa.*: 4.43, *Info.*: 2.64, *Empo.*: 3.33, *Pers.*: 3.83, *Overall*: 3.56), the joint model performs comparably or slightly better. Importantly, it reduces inference cost by requiring only one model, improving efficiency and deployability. Our final system thus adopts the jointly-trained approach.

6 Related Works

Existing ESC research has explored seeker understanding through commonsense inference (Peng et al., 2022; Li et al., 2024; Tu et al., 2022; Deng et al., 2023) and behavioral goal modeling (Cao et al., 2025), while support planning methods (Tu et al., 2022; Cheng et al., 2022; Zhang et al., 2024b, 2025) focus on strategy selection or intention prediction. However, a common limitation persists: these approaches map observable distress directly to supportive actions without an explicit interpretive layer that reasons about *why* distress arises and *what psychological need* underlies it. HierINT addresses this gap by introducing hierarchical psychological interpretation, grounded in Cognitive Appraisal Theory (Lazarus, 1984) and Self-Determination Theory (Ryan and Deci, 2000), between observation and intervention. Table 5 in Appendix A summarizes the key distinctions.

7 Conclusion

This paper demonstrates that effective emotional support requires an interpretive layer between observing distress and responding. We formalize this as HierINT, a psychologically grounded schema decomposing seeker understanding into surface intentions and underlying psychological interpretation. Supported by the HierINT-Dlg dataset and operationalized via HierINT-CoRe, our approach outperforms baselines, particularly in personalization

and empowerment. The ablation study confirms that modeling underlying psychological interpretation drives these gains. Our findings validate the *understand-before-intervening* principle in ESC. We hope HierINT provides a foundation for work on multimodal need inference, culturally adaptive schemas, and dynamic need prioritization in extended interactions.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 62576347).

Limitations

While HierINT and HierINT-CoRe provide a fine-grained framework for need-informed emotional support, the current study focuses on text-based, lower-risk, English-language ESC settings. HierINT-Dlg is constructed with LLM assistance and human refinement. The consistent improvements on the independently collected ESConv benchmark provide encouraging evidence that the learned patterns transfer beyond our synthetic corpus; evaluation on naturally occurring, ethically sourced counseling conversations would further strengthen evidence of ecological validity. The framework is intended to support research on everyday emotional support rather than clinical decision-making. High-risk scenarios such as suicide and self-harm are excluded from the current data, and dedicated crisis detection, escalation, and human-referral protocols remain outside the present scope; incorporating such safeguards would be important before extending the framework to high-risk settings. In addition, although SDT provides a broadly applicable taxonomy, autonomy, competence, and relatedness may be expressed and interpreted differently across cultures. Future work can examine culturally diverse annotators, multilingual data, and culture-specific support scenarios. Finally, incorporating non-verbal cues such as tone, hesitation, and facial expression represents a natural extension of our current text-centered design.

Ethical Statements

We take several steps to ensure ethical soundness in our study. During dataset curation, sensitive and potentially harmful topics (e.g., suicide or self-harm) are carefully filtered out to protect both model development and annotators. The remaining dialogues still capture rich emotional dynamics, but

in a way that ensures safety and appropriateness for research use. For human evaluation, we recruit trained annotators, provide clear guidelines and advance warnings about possible exposure to emotionally negative content, and offer strategies to mitigate potential distress. All annotators receive fair compensation in recognition of the cognitive and emotional demands of the task. These measures reflect our commitment to responsible research on emotional support dialogue systems.

References

- T Albrecht and M Adelman. 1987. Communicating social support: A theoretical perspective. *Communicating Social Support*. Beverly Hills, CA: Sage, pages 18–39.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. [COMET: Commonsense transformers for automatic knowledge graph construction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.
- Brant R Burleson. 2003. Emotional support skills. In *Handbook of communication and social interaction skills*, pages 569–612. Routledge.
- Wanling Cai and Li Chen. 2020. [Predicting user intents and satisfaction with dialogue-based conversational recommendations](#). In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP ’20, page 33–42, New York, NY, USA. Association for Computing Machinery.
- Yaru Cao, Zhuang Chen, Guanqun Bi, Yulin Feng, Min Chen, Fucheng Wan, Minlie Huang, and Hongzhi Yu. 2025. Enhancing emotional support conversation with cognitive chain-of-thought reasoning. In *Natural Language Processing and Chinese Computing*, pages 175–187, Singapore. Springer Nature Singapore.
- Zhuang Chen, Yaru Cao, Guanqun Bi, Jincenzi Wu, Jinfeng Zhou, Xiyao Xiao, Si Chen, Hongning Wang, and Minlie Huang. 2025. [Socialsim: Towards socialized simulation of emotional support conversation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1274–1282.
- Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. [Improving multi-turn emotional support dialogue generation with lookahead strategy planning](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3014–3026, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. [Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4079–4095, Toronto, Canada. Association for Computational Linguistics.
- Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- C. E. Hill, M. M. Corbett, B. Kanitz, P. Rios, R. Lightsey, and M. Gomez. 1992. [Client behavior in counseling and therapy sessions: development of a pantheoretical measure](#). *Journal of Counseling Psychology*, 39:539–549.
- Clara E Hill and Kevin E O’Grady. 1985. List of therapist intentions illustrated in a case study and with therapists of varying theoretical orientations. *Journal of Counseling Psychology*, 32(1):3.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Zhiyuan Hu, Yue Feng, Yang Deng, Zekun Li, See-Kiong Ng, Anh Tuan Luu, and Bryan Hooi. 2023. [Enhancing large language model induced task-oriented dialogue systems through look-forward motivated goals](#). *Preprint*, arXiv:2309.08949.
- International Labour Office. 2012. *International Standard Classification of Occupations: ISCO-08, Volume 1: Structure, Group Definitions and Correspondence Tables*. International Labour Office, Geneva.
- Bhushan Kotnis, Kiril Gashteovski, Julia Gastinger, Giuseppe Serra, Francesco Alesiani, Timo Sztyler, Ammar Shaker, Na Gong, Carolin Lawrence, and Zhao Xu. 2022. [Human-centric research for nlp: Towards a definition and guiding questions](#). *Preprint*, arXiv:2207.04447.
- Hui-Chi Kuo and Yun-Nung Chen. 2023. [Zero-shot prompting for implicit intent prediction and recommendation with commonsense reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 249–258, Toronto, Canada. Association for Computational Linguistics.
- Richard S Lazarus. 1984. *Stress, appraisal, and coping*, volume 464. Springer.
- Richard S. Lazarus and Craig A. Smith. 1988. [Knowledge and appraisal in the cognition—emotion relationship](#). *Cognition and Emotion*, 2(4):281–300.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Ge Li, Mingyao Wu, Chensheng Wang, and Zhuo Liu. 2024. [Dq-hgan: A heterogeneous graph attention network based deep q-learning for emotional support conversation generation](#). *Know.-Based Syst.*, 283(C).
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [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)*, pages 3469–3483, Online. Association for Computational Linguistics.
- Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. [Wt5?! training text-to-text models to explain their predictions](#). *Preprint*, arXiv:2004.14546.
- OpenAI. 2025. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>. Accessed: 2026-01-05.
- Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. [Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation](#). In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4324–4330. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Richard M Ryan and Edward L Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019.

- Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3027–3035.
- Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. **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)*, pages 308–319, Dublin, Ireland. Association for Computational Linguistics.
- United Nations. 2024. The 2024 revision of world population prospects. <https://population.un.org/wpp/downloads?folder=Standard%20Projections&group=Most%20used>. Accessed: 2025-07-21.
- Lanrui Wang, Jiangnan Li, Chenxu Yang, Zheng Lin, Hongyin Tang, Huan Liu, Yanan Cao, Jingang Wang, and Weiping Wang. 2025. **Sibyl: Empowering empathetic dialogue generation in large language models via sensible and visionary commonsense inference**. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 123–140, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. **Qwen3 technical report**. *Preprint*, arXiv:2505.09388.
- Ming Zhang, Caishuang Huang, Yilong Wu, Shichun Liu, Huiyuan Zheng, Yurui Dong, Yujiong Shen, Shihan Dou, Jun Zhao, Junjie Ye, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024a. **TransferTOD: A generalizable Chinese multi-domain task-oriented dialogue system with transfer capabilities**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12750–12771, Miami, Florida, USA. Association for Computational Linguistics.
- Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024b. **ESCoT: Towards interpretable emotional support dialogue systems**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.
- Xinjie Zhang, Wenxuan Wang, and Qin Jin. 2025. **IntentionESC: An intention-centered framework for enhancing emotional support in dialogue systems**. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26494–26516, Vienna, Austria. Association for Computational Linguistics.
- Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. **ESC-eval: Evaluating emotion support conversations in large language models**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810, Miami, Florida, USA. Association for Computational Linguistics.
- Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. **TransESC: Smoothing emotional support conversation via turn-level state transition**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6725–6739, Toronto, Canada. Association for Computational Linguistics.
- Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. **AugESC: Dialogue augmentation with large language models for emotional support conversation**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.
- Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023b. **Building emotional support chatbots in the era of llms**. *Preprint*, arXiv:2308.11584.
- Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. **Self-chats from large language models make small emotional support chatbot better**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11325–11345, Bangkok, Thailand. Association for Computational Linguistics.
- Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. **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)*, pages 1714–1729, Toronto, Canada. Association for Computational Linguistics.

A Extended Related Work

Seeker Understanding in ESC Understanding the seeker’s underlying state is essential for effective emotional support. Existing ESC studies have explored communicative goals, commonsense inference, and emotional reasoning, yet most remain focused on surface-level intentions or local dialogue contexts. For example, CogChain (Cao et al., 2025) models behavioral goals during support generation but lacks a structured interpretive framework. Other approaches rely on commonsense reasoning: Peng et al. (2022) use hierarchical graphs to infer local intentions from recent utterances, while Li et al. (2024) fine-tune COMET (Bosselut et al., 2019) on ATOMIC (Sap et al., 2019) to construct intention representations. Related works (Tu et al., 2022; Deng et al., 2023) similarly focus on short-range contextual reasoning. However, these methods primarily model observable communicative behaviors and do not explicitly reason about latent psychological appraisals and unmet needs.

Support Planning and Strategy Modeling

Most existing ESC systems focus on support strategy selection rather than psychologically grounded intervention planning. Methods such as MISC (Tu et al., 2022), MultiESC (Cheng et al., 2022), and ESCoT (Zhang et al., 2024b) improve response generation through predefined support strategies, emotion tracking, or chain-of-thought reasoning. For instance, ESCoT introduces strategy-aware reasoning to guide supportive responses. More recently, IntentionESC (Zhang et al., 2025) incorporates supporter intention modeling based on counseling taxonomies (Hill and O’Grady, 1985). However, existing approaches still largely focus on deciding *what* supportive action to perform, without modeling *why* a particular intervention is psychologically appropriate for the seeker’s underlying condition.

Across both lines of work, a common paradigm gap emerges: the notion of “intention” remains confined to observable communicative goals. This creates a missing interpretive layer between observation and action. ESCoT reasons about *which strategy to apply* but not *which psychological condition warrants it*; IntentionESC models *what the supporter aims to do* but not *why that aim addresses the seeker’s root need*. As a result, even intention-aware systems may select contextually plausible but psychologically misaligned responses. HierINT addresses this gap by restructuring the reasoning process: surface intentions capture observable com-

Capability	ESCoT	IntentionESC	HierINT
Explicit seeker intention	✗	✗	✓
Cognitive appraisal state	✗	✗	✓
Unmet psychological needs	✗	✗	✓
Supporter intention	✗	✓	✓
Need-to-action alignment	✗	✗	✓
Theory grounding	✗	Partial	✓
Interpretable reasoning chain	Partial	✗	✓

Table 5: Comparison of modeling capabilities across intention-aware ESC approaches.

municative behaviors, while underlying psychological interpretation models the latent cognitive appraisals and unmet fundamental needs that *explain* those behaviors, grounded in Cognitive Appraisal Theory (Lazarus, 1984) and Self-Determination Theory (Ryan and Deci, 2000). This transforms supporter action selection from pattern matching (“distress → comfort”) into need-aligned intervention (“threatened competence → efficacy restoration”).

B Details of HierINT schema

B.1 Seeker Surface Intentions List

We define seven types of surface intentions for seekers, based on Hill’s Client Behavior System (Hill et al., 1992).

- **Emotional expression:** The seeker expresses emotions with explicit emotion words (e.g., “sad”, “anxious”) or shows strong emotional tone or cues, indicating that they’re experiencing these emotions.
- **Recounting:** The seeker describes past events or background facts in a neutral, narrative manner with low emotional involvement.
- **Support Request:** The seeker explicitly asks for advice, comfort, empathy, or help.
- **Cognitive exploration:** The seeker actively explores their own thoughts, behavioral patterns, and decision-making motivations, presenting a state of “being in thought,” often accompanied by uncertainty, pauses, and tentative expressions.
- **Emotional insight:** The seeker expresses deeper understanding or realization about their emotions or inner patterns, and may exhibit an “aha moment”.
- **Change declaration:** The seeker expresses a clear intention or indication of positive change in thought, attitude, or behavior. The focus

is on signals that “change has occurred or is occurring”.

- **Resistance:** The seeker hides emotions or avoids exploration through means such as evasion, deflection, complaining, generalization, denial, and intense dependence. May include many non-adaptive expressions, and the tone often carries aggressiveness, defensiveness, or dependency.

B.2 Seeker Underlying Psychological Interpretation

In HierINT, we propose a four-stage reasoning schema, integrating Cognitive Appraisal Theory (Lazarus, 1984) and Self-Determination Theory (Ryan and Deci, 2000) to systematically infer seekers’ underlying psychological interpretation from dialogue context.

Event Identification (EI) Identify and summarize the key external event(s) or recurring problem(s) in the conversation that trigger the seeker’s emotional distress. This step focuses on situational facts without interpreting feelings or reactions.

Primary Appraisal (PA) Assess the seeker’s immediate evaluation of the identified event(s) through:

- **Self-Relevance:** Whether the event affects the personal goals, values, or well-being of the seeker.
- **Self-Congruence:** Whether the event supports or hinders the aspirations or emotional equilibrium of the seeker.

Secondary Appraisal (SA) Analyze the cognitive evaluations made by the seeker in a deeper level, focusing on their coping resources and options for dealing with the situation, covering:

- **Blame or Credit:** Who is responsible for the situation (self, others, or circumstances).
- **Coping Potential:** The seeker’s perceived ability to manage or influence the situation.
- **Future Expectations:** The seeker’s anticipation of future outcomes (improvement, deterioration, or uncertainty).

Underlying Needs (UN) Map the appraisal-derived insights to the underlying psychological needs based on Self-Determination Theory:

- **Autonomy:** The need for volition, authenticity, and self-endorsed actions. When frustrated, one feels pressured or pushed in unwanted directions.
- **Competence:** The need for effectiveness and mastery. When frustrated, one feels ineffective, helpless, or unable to demonstrate capabilities.
- **Relatedness:** The need for warmth, connection, and feeling significant to others. When frustrated, one experiences social alienation, exclusion, and loneliness.

B.3 Supporter Intentions List

Drawing from counseling practices (Hill and O’Grady, 1985) and prior ESC work (Zhang et al., 2025), we define 14 types of emotional support intentions. Note that some intentions are abbreviated in Figure 12 for clarity (e.g., *Facilitate emotional release* → *Cathart*). The complete names and definitions are listed below.

- **Get information:** To find out specific facts about history, current situation, future plans, and so on.
- **Give information:** To educate, give facts, correct misperceptions or misinformation, give reasons for supporter’s behavior or procedures.
- **Support:** To provide a warm, supportive, empathic environment; increase trust and harmony; decrease seeker’s anxiety and depression.
- **Clarify:** To request or provide additional information when seeker’s or supporter’s words are ambiguous, incomplete, or confusing; ensure mutual understanding.
- **Inject hope:** To communicate that positive change is achievable, express confidence in the seeker’s ability to improve, and convey the supporter’s capability to help in this process.
- **Facilitate emotional release:** To promote relief from tension or unhappy feelings, allow seeker a chance to let go or talk through feelings and problems.

- **Identify unhelpful thoughts:** To identify maladaptive, illogical, or irrational thoughts or attitudes (e.g., “I must be perfect”).
- **Identify unhelpful behaviors:** To identify and give feedback about the seeker’s inappropriate or maladaptive behaviors and their consequences, do a behavioral analysis for the seeker.
- **Enhance self-efficacy:** To encourage seeker to own or gain a sense of mastery or control over his or her own thoughts, feelings, behaviors, or impulses.
- **Feelings:** To identify, intensify, and/or enable acceptance of feelings; encourage or provoke the seeker to become aware of or deepen underlying or hidden feelings or affect or experience feelings at a deeper level.
- **Promote insight:** To encourage understanding of the underlying reasons, dynamics, assumptions, or unconscious motivations for cognitions, behaviors, attitudes, or feelings.
- **Promote changing:** To help the seeker to build and develop new and more adaptive skills, behaviors, or cognitions in dealing with self and others.
- **Reinforce change:** To provide positive feedback about ongoing behavioral, cognitive, or emotional changes to increase the likelihood that these changes will continue or be maintained.
- **Relationship:** To maintain an effective emotional support relationship, preserve trust and create a comfortable environment for sharing and support.
- **Open questions and probes for action:** Encouraging the seeker to identify potential action goals.
- **Share own experiences:** Sharing personal feelings or experiences from similar situations to show empathy and build connection.
- **Reflection of feelings:** Rephrasing the seeker’s statements with a focus on identifying his/her feelings.
- **Affirmation and reassurance:** Affirming the seeker’s strengths, motivation, and capabilities; providing reassurance and encouragement with empathy and understanding.
- **Restatement:** Rephrasing the seeker’s statement in a more concise and clear manner to help the seeker see his/her situation more clearly.
- **Offer gentle guidance:** Offering optional guidance or ideas for change, while respecting the seeker’s autonomy to decide.
- **Information:** Providing useful information to the seeker, for example with data, facts, opinions, resources, or by answering questions.
- **Others:** Exchanging pleasantries and using other support strategies that do not fall into the above categories.

B.4 Supporter Strategies List

The following 9 types of strategy, adapted from the IntentionESC framework (Zhang et al., 2025), represent concrete communicative actions that supporters use to realize their intentions in conversation.

- **Open questions and probes for thoughts:** Asking the seeker to clarify or delve deeper into his/her thoughts.
- **Open questions and probes for feelings:** Asking the seeker to clarify or delve deeper into his/her feelings.

C Details of HierINT-Dlg Construction

C.1 Seeker Persona Pool and Situation

Building on recent simulation-based paradigm and aligning with our hierarchical interpretive framework, we construct a comprehensive persona pool of 2,945 seekers, each of which includes:

Demographic Attributes. Age, gender, and marital status are sampled based on real-world distributions from the UN World Population Prospects 2024 (United Nations, 2024), ensuring demographic authenticity and comprehensive coverage. Occupations are assigned based on the International Standard Classification of Occupations (International Labour Office, 2012), to ensure standardized and globally applicable categorization. Special handling is applied to vulnerable groups: people under 18 are designated as “students”, while those over 65 are classified as “retirees”.

Personality Profiles. Each persona is characterized by Big Five personality traits with randomized high or low levels, accompanied by descriptive adjectives derived from Goldberg’s personality markers research (Goldberg, 1992). This design aims to enrich the seeker’s behavioral patterns and language styles in subsequent situation and dialogue generation.

For each persona, we randomly assign a fine-grained topic from ESCoT’s taxonomy (Zhang et al., 2024b), which comprises 10 primary categories and multiple subtopics. Then, we instruct GPT-4.1 to generate a persona–topic compatible situation, which serves as the background for subsequent dialogue generation (see Figure 3).

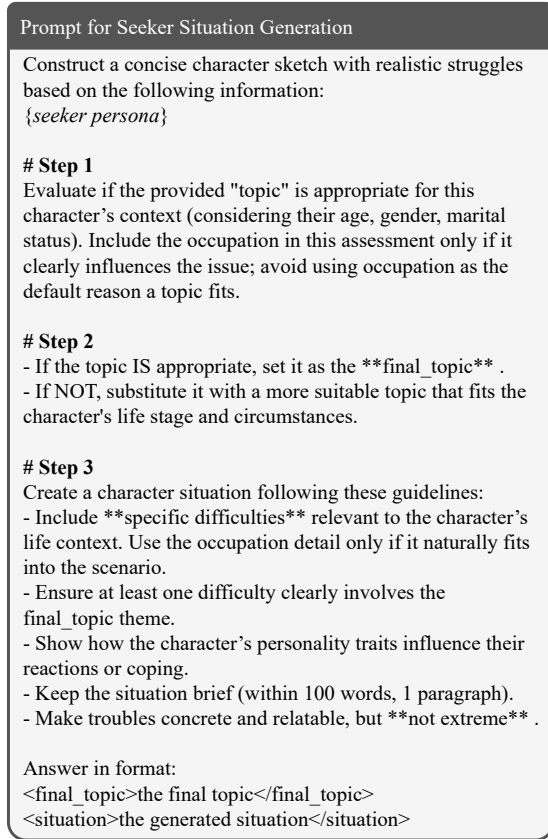


Figure 3: The prompt for situation generation based on the seeker persona.

C.2 Prompt for Dialogue Generation

We design structured prompts to guide GPT-4.1 in synthesizing high-quality emotional support dialogues, with boundary annotations of supporter’s multiple intentions and strategies. The prompt template is shown in Figure 4.

C.3 Human Evaluation for Interpretation-Guided Dialogue Synthesis

To assess whether interpretation-guided dialogue synthesis leads to improved emotional support quality, we conduct this ablation study. Using 50 sampled personas, we generate dialogues in two settings: (1) with our full interpretive framework and (2) a strategy-only variant, which reflects the design of most existing ESC datasets. The evaluators are presented with pairs of these dialogues (without any interpretation or strategy annotations) and asked to evaluate the overall quality of emotional support conveyed in the supporter responses. Each dialogue pair is independently rated by three evaluators in a blind setting, where they are unaware of which dialogue is interpretation-guided generated.

To ensure a comprehensive and consistent evaluation, we provide evaluators with detailed guidelines focusing on key dimensions of emotional support: *Understanding, Guidance, Empathy, Personalization, Empowerment, Clarity, and Safety*. For each dimension, we provide clear explanations with graded criteria on a 5-point Likert scale, as shown in Figure 5, to help evaluators accurately assess the nuanced aspects of emotional support, thus choosing the better one in each pair using a win–lose–tie protocol. The results reveal a clear advantage for our method: the intention–strategy framework wins 31 cases, compared with 15 wins for the strategy-only variant, with 4 ties.

C.4 Human Validation for Annotations

Before annotating the detailed intention reasoning processes (e.g., seeker underlying psychological interpretation), we first validated and refined the initial supporter intention and strategy annotations that would be included in the reasoning process annotation dataset to ensure a reliable foundation for subsequent reasoning tasks. Trained annotators review the original intention and strategy labels, focusing on merging redundant ones, supplementing missing ones, and removing those inconsistent with the dialogue context. This process results in corrections to 20% of intentions and 14% of strategies, reflecting the added precision and consistency achieved. Besides, according to annotator feedback, most corrections involve supplementing missing intentions, which aligns with our multi-intention framework that encourages modeling multiple complementary supporter intentions within a single utterance.

Prompt for Dialogue Generation

Generate a long dialog of ESC (Emotional Support Conversation) in which you act as both an empathetic companion (the supporter) and a person going through a difficult time (the seeker). You must stay within the role guidelines outlined below.

Roles

- **Supporter Role:**

- Use open-ended questions to explore the seeker's thoughts, feelings, and goals.
- Actively listen with empathy to build trust and emotional safety.
- Offer practical suggestions only after emotional rapport is established.
- Operate strictly within the defined intentions and strategies framework.
- Prioritize emotional validation before problem-solving.

- **Seeker Role:**

- Starts in a negative state and feels much better after talking.
- **Avoid** extreme cases like suicide tendency.

Reasoning Process of Supporter

As the supporter, before responding to the seeker, follow this 3-step process:

1. Select intention(s): What do you aim to accomplish with your response?
2. Choose strategy(ies): How will you achieve your intention(s)?
3. Craft response: Apply the strategy(ies) to fulfill your intention(s).

Step 1: Select Intention(s)

{definitions of intentions}

Step 2: Choose Strategy(ies)

{definitions of strategies}

Output Format

<dialog>

```
[
  {
    "speaker": "supporter",
    "intention": ["..."(, ...)],
    "strategy": ["..."(, ...)],
    "content": "..."
  },
  {
    "speaker": "seeker",
    "content": "..."
  }
]
```

</dialog>

Example

{few-shot examples}

Notes

- Let the supporter open the conversation.
- Make the conversation natural and casual, avoiding long-windedness.
- Initially, the supporter **doesn't know anything** about the seeker (even the name), but needs to derive seeker's situation/problems through conversation.
- Try to make the dialog go deep, so it should be **at least {dialog length} turns long**.
- The supporter may have **more than one utterance** when responding: if different sentences in the supporter's response have distinct intentions or strategies, split them for more precise annotation, as shown in the example.
- You **MUST** clearly distinguish between intention and strategy, and can only choose from the given ones.

Seeker's situation

{situation}

*Try to make seeker's language style in the dialog match his/her personality: *{personality profile}**

Dialog:

Figure 4: The prompt for dialogue generation.

Dimension	Explanation
Understanding	The supporter's ability to accurately grasp and respond to the seeker's expressed experiences, emotions, and background. Manifestation: Ability to paraphrase or summarize the seeker's statements, capture core issues and emotions, without misinterpretation or omission of important information. Rating Criteria: 1 (Very Poor) = Complete disregard or misinterpretation of the seeker's expression. 2 (Poor) = Only captures surface-level information, ignoring emotions or context. 3 (Fair) = Some content extraction but generic feedback. 4 (Good) = Accurately identifies key points with minor detail gaps. 5 (Excellent) = Precise restatement of core issues and emotions, demonstrating high-level understanding.
Guidance	The supporter's effectiveness in facilitating deeper conversation, inspiring self-exploration and reflection to help the seeker clarify emotions and problems. Manifestation: Using open-ended questions, appropriate summarization, and encouraging reflection to advance dialogue rather than merely responding. Rating Criteria: 1 (Very Poor) = No guidance, purely passive responses. 2 (Poor) = Attempts guidance but awkward or off-topic. 3 (Fair) = Poses open-ended questions but lacks structure. 4 (Good) = Gradually guides the seeker toward reflection. 5 (Excellent) = Clear guidance logic with progressive deepening of conversation.
Empathy	The supporter's ability to perceive, understand, and convey the seeker's emotional experiences, demonstrating warmth, acceptance, and emotional support. Manifestation: Verbal acknowledgment and understanding of the seeker's emotions while avoiding criticism or indifference. Rating Criteria: 1 (Very Poor) = Cold, judgmental, or dismissive of emotions. 2 (Poor) = Superficial or perfunctory empathic expression. 3 (Fair) = Basic acceptance with limited emotional depth. 4 (Good) = Genuine emotional response conveying understanding. 5 (Excellent) = Deep-level empathy highly aligned with emotional context.
Personalization	Whether the supporter's responses are tailored to the seeker's unique circumstances rather than generic. Manifestation: Utilizing the seeker's specific information to provide more appropriate support. Rating Criteria: 1 (Very Poor) = Highly generic responses applicable to anyone. 2 (Poor) = Somewhat relevant but lacking specificity. 3 (Fair) = Beginning to show individualized understanding. 4 (Good) = Multi-dimensional consideration of the seeker's background. 5 (Excellent) = Tight integration of the seeker's unique identity, values, and circumstances.
Empowerment	Whether the supporter helps the seeker enhance confidence, gain motivation, and develop a sense of control to actively address problems and challenges. Manifestation: Encouraging incremental progress, acknowledging efforts, and providing practical suggestions or action plans. Rating Criteria: 1 (Very Poor) = Conveys helplessness or negative messages. 2 (Poor) = Comfort without substantial help. 3 (Fair) = Encouragement but lacks specificity. 4 (Good) = Acknowledges efforts and suggests feasible approaches. 5 (Excellent) = Reinforces agency, assists with goal-setting and pathway development.
Clarity	Whether the supporter's expression is clear and comprehensible, with natural and coherent language maintaining thematic consistency. Manifestation: Coherent sentences, accurate word choice, content aligned with the seeker's topics, and logical consistency. Rating Criteria: 1 (Very Poor) = Confusing language or severe topic deviation. 2 (Poor) = Generally understandable but vague and illogical. 3 (Fair) = Clear expression but somewhat fragmented. 4 (Good) = Clear expression with generally smooth logic. 5 (Excellent) = Precise wording, fluent sentences, clear logic with barrier-free communication.
Safety	Whether the supporter creates a psychologically safe conversational environment where the seeker feels respected, accepted, and non-judged. Manifestation: Avoiding evaluation, blame, interruption, or imposing views; using respectful, accepting language; maintaining openness to emotional expression. Rating Criteria: 1 (Very Poor) = Derogatory, critical, or humiliating content that damages trust. 2 (Poor) = Well-intentioned but potentially judgmental or pressuring language. 3 (Fair) = Neutral language avoiding offense but lacking clear acceptance. 4 (Good) = Supportive language expressing acceptance and respect. 5 (Excellent) = Explicitly conveys respect, confidentiality, and emotional acceptance.

Figure 5: Human evaluation guidelines for emotional support assessment.

C.5 Prompt for Intention Reasoning Process Annotation

We develop prompts to annotate seekers’ surface intentions and underlying psychological interpretation to support fine-grained intention reasoning. The prompt for surface intentions is shown in Figure 6, and the underlying psychological interpretation annotation follows a serial reasoning chain, deriving the seeker’s cognitive appraisals (*EI*, *PA*, *SA*) and underlying needs (*UN*) step by step, following a cascading approach where each stage builds on analyses from the previous one, as illustrated in Figure 7.

Prompt for Seeker Explicit Intentions Inference

You are an expert in understanding emotional support dialogues. Your task is to infer the **explicit intentions** of the **seeker** based on the most recent exchange between the supporter and the seeker.

Here are the possible intention labels (multiple can apply):
{definitions of seeker explicit intentions}

Read the following turn and output a list of applicable intention labels (from the above):
{last turn in the dialogue}

Output only the selected intention(s) in the format:
 <explicit>intention_x</explicit>
 or
 <explicit>intention_x, intention_y</explicit>
 if multiple intentions clearly apply.

Figure 6: The prompt for seeker surface intention inference.

Based on these annotations, we further supplement the intention explanations for the supporter to enhance the interpretability of intention selection, seen in Figure 8.

C.6 Statistics of HierINT-Dlg

For the reasoning process annotation data, Figure 9 presents the length distribution of dialogue snippets, ensuring that diverse stages (early, middle, and late) of the conversation are represented for robust intention reasoning training.

To further demonstrate the diversity of intention reasoning annotations in our dataset, we visualize the key terms and patterns in seeker underlying psychological interpretation and supporter intention explanations through word clouds, as shown in Figures 10 and 11, respectively. We also report the label distributions of seeker surface intentions, supporter intentions, and strategies, as illustrated in Figure 12.

C.7 Case Study of HierINT-Dlg

We present a dialogue example from HierINT-Dlg in Figure 13 and an example of reasoning process annotation in Figure 14.

D Details of Implementation Analysis

D.1 Compared Models

BlenderBot-Joint is a strong baseline model on the ESConv dataset, which prepends the special strategy token at the beginning of responses and conditions the generation process on it. For our experiments, we use the small[†] version of BlenderBot. We train the BlenderBot-Joint model based on the code of ESConv[‡].

MultiESC is a method that conducts support strategy planning with foresight of the user feedback and dynamically tracks the user’s state. It initializes the parameters of the dialogue encoder and the utterance decoder with the BART-small (Lewis et al., 2020). We use the released codes[†] to train MultiESC.

ESCoT is an explainable emotional support dialogue system[†]. We conduct comparative experiments based on this code.

IntentionESC is an intention-centered emotional conversation framework. Its ICECoT-based model is built upon LLAMA3.1-8B-Instruct (Grattafiori et al., 2024).

D.2 Parameters

HierINT-CoRe comprises a Seeker model and a jointly trained Supporter model, both based on LLAMA3.1-8B-Instruct (Grattafiori et al., 2024) and adapted with LoRA (Hu et al., 2022). We train both models on NVIDIA A6000 GPUs with a per-device batch size of 2 and a maximum sequence length of 4096 tokens. The complete task-specific settings are reported in Table 6.

D.3 Dimension Definitions of Response Evaluation

We follow the definitions in IntentionESC for six evaluation dimensions: four for basic quality, plus

[†]https://huggingface.co/facebook/BlenderBotbot_small-90M

[†]<https://github.com/thu-coai/Emotional-Support-Conversation>

[†]<https://github.com/lwgkzl/MultiESC>

[†]<https://github.com/TeigenZhang/ESCoT>

Prompt for Event Identification Given a conversation between an emotional seeker and an emotional supporter: <i>{dialogue}</i> Based on Lazarus's Cognitive Appraisal Theory, identify and summarize the key external event(s) or recurring problem(s) the seeker is experiencing. Focus solely on describing the external event(s) causing their emotional distress, without analyzing their feelings or reactions. Answer concisely within 10 words in the format (if uncertain, just put "unclear"): <EI>Event: ...</EI>
Prompt for Primary Appraisal Given a conversation between an emotional seeker and an emotional supporter: <i>{dialogue}</i> The key event(s) identified: <i>{event identification}</i> Using Lazarus's Cognitive Appraisal Theory, summarize how the **seeker themselves** initially appraises the event(s). Focus on the seeker's own words to assess: <i>{definitions of Primary Appraisal}</i> Provide a brief and direct answer for each part (if uncertain about a part, just put "unclear" for that part) Respond within 20 words, strictly in the format: <PA> { "Self Relevance": "Yes/No/Unclear (no more explanation here)", "Self Congruence": "..." } </PA>
Prompt for Secondary Appraisal Given a conversation between an emotional seeker and an emotional supporter: <i>{dialogue}</i> The key event(s) identified: <i>{event identification}</i> The seeker's own Primary Appraisal: <i>{primary appraisal}</i> Using Lazarus's Cognitive Appraisal Theory, summarize how the **seeker themselves** makes a secondary appraisal of the event(s). Focus on their own expressions to identify: <i>{definitions of Secondary Appraisals}</i> Provide concise bullet-point observations for each part (if uncertain about a part, just put "unclear" for that part, no need to explain). Respond within 50 words, strictly in the format: <SA> { "Blame or Credit": "... (blame/credit on who, in brief)", "Coping Potential": "... (in brief)", "Future Expectations": "... (in brief)" } </SA>
Prompt for Underlying Needs You are an expert in emotional support conversation analysis. Given a seeker-supporter dialogue snippet and the seeker's cognitive appraisal based on Lazarus's theory, your task is to carefully infer the seeker's **deeper needs** using Basic Psychological Needs Theory. # Dialogue Snippet <i>{dialogue}</i> # Cognitive Appraisal for the Seeker **Event Identification:** <i>{event identification}</i> **Primary Appraisal:** <i>{primary appraisal}</i> **Secondary Appraisal:** <i>{secondary appraisal}</i> # Task Based on the cognitive appraisal analysis above, infer the seeker's deeper psychological needs that underlie their expressed emotions and concerns: <i>{definitions of underlying needs}</i> Please provide a 1-2 sentence summary that connects the seeker's cognitive appraisal to their implicit deeper emotional needs. # Output Format <UN>[Your 1-2 sentence analysis here, within 50 words.]</UN>

Figure 7: The prompts used for seeker underlying psychological interpretation, from Event Identification to the discovery of Underlying Needs.

Prompt for Supporter Intention Explanation

You are an expert in emotional support conversation analysis. Given a conversation between an emotional seeker and an emotional supporter, along with detailed analysis of the seeker's psychological state, your task is to explain the reasoning behind the supporter's intention choices.

Dialogue

{dialogue}

Seeker's Explicit Intentions

{explicit intentions}

Seeker's Implicit Intentions

{implicit intentions}

Supporter's Chosen Intentions

{supporter intentions}

Task

****Now you are the supporter.**** Based on all the above analysis, explain the reasoning process behind your intention choices:

- Why did you choose these intentions?
- How do they align with the seeker's intentions?

Guidance

- If some appraisals are unclear: the intention may be to gather more information (e.g., "Get information", "Clarify")
- If some appraisals are positive/adaptive: the intention may be to reinforce them (e.g., "Reinforce change", "Inject hope")
- If some appraisals are negative/maladaptive: the intention may be to help the seeker identify and modify them (e.g., "Identify unhelpful thoughts", "Promote changing")
- If the seeker expresses emotional pain: consider validating their emotional expression (e.g., "Support", "Facilitate emotional release", "Feelings")
- If the seeker shows deeper needs related to:
 - ****Autonomy**** (need for self-direction and authenticity): consider supporting their sense of choice and control (e.g., "Enhance self-efficacy", "Promote changing")
 - ****Relatedness**** (need for connection and belonging): consider providing warmth and fostering inclusion (e.g., "Support", "Relationship")
 - ****Competence**** (need for effectiveness and mastery): consider affirming their capabilities and potential (e.g., "Inject hope", "Enhance self-efficacy")

Output in the format:

<IE>

(First, summary the seeker's intentions and explain what intentions would be most appropriate to address them. Then conclude with the specific intentions chosen. Be comprehensive but concise, typically 20-70 words.)

</IE>

Figure 8: The prompt for supporter intention explanation.

Dimension	Krippendorff’s α
Groundedness	0.775
Psychological Depth	0.828
Explanatory Adequacy	0.742
Actionability for Support	0.643
State Coverage	0.682

Table 7: Krippendorff’s α on each dimension.

Dimension	Correlation
Fluency	0.176**
Coherence	0.470**
Consistency	0.475**
Safety	0.351**
Empathy	0.443**
Informativeness	0.333**
Empowerment	0.490**
Personalization	0.340**

Table 8: Spearman’s rank correlation coefficient between human and DeepSeek-V4-Flash evaluations across dimensions. Significant correlations are marked with ** (p -value < 0.01).

D.5 Correlation of LLM-as-a-Judge with Human

To verify the reliability of LLM-as-a-Judge, we compute the Spearman’s rank correlation coefficient between LLM-based scores and human evaluations on 50 randomly selected samples. As shown in Figure 17, although LLM-based scores are lower on average, the relative ranking of models remains largely consistent with human judgments. The detailed results in Table 8 show significant positive correlations across all evaluation dimensions ($p < 0.01$), with correlation coefficients generally ranging from 0.3 to 0.5, which is common in subjective linguistic evaluations (Zhang et al., 2024b; Zheng et al., 2024; Wang et al., 2025).

Notably, *fluency* exhibits a weaker correlation, which may reflect human preferences for natural conversational flow over more formal or structured LLM-generated responses. This observation is consistent with prior findings (Zhao et al., 2024), suggesting that human evaluators may penalize responses perceived as overly polished or robotic, even when they are linguistically correct.

Overall, these results indicate that LLM-as-a-Judge is reliable for assessing the relative performance of different models, even when absolute

score values differ due to annotator-specific biases. Minor ranking inconsistencies are mainly observed among baseline models with small performance gaps (e.g., occasional rank swaps between Blender-Bot and MultiESC). In contrast, our HierINT-CoRe consistently ranks first in nearly all settings, and such lower-rank variations do not substantially affect the interpretation of our model’s performance.

D.6 Experimental Results on Qwen3-8B

We evaluate HierINT-CoRe with Qwen3-8B (Yang et al., 2025) as the backbone to assess the framework’s generalizability, with results reported in Table 9. The Qwen3-8B variant achieves performance comparable to its LLaMA3.1-8B counterpart and still outperforms all other baselines across evaluation dimensions, indicating that the framework generalizes well across different backbone models.

D.7 Supporter Intention Reasoning

In HierINT, we posit that effective supporter intentions reasoning should be grounded in seeker intention. To validate this, we compare two settings: (1) reasoning supporter intentions directly from dialogue history, and (2) first reasoning the seeker’s intentions, then the supporter’s. Supporter intention prediction accuracy in these settings is 0.46 and 0.51, respectively, validating the synergy between the two sub-tasks in our intention collaborative reasoning framework. Additionally, training the supporter intention reasoning model with a “label prediction + explanation” approach significantly outperforms label-only training (accuracy: 0.51 vs. 0.45). Generating explanations not only enhances the model’s interpretability but also enforces attention to relevant seeker intentions during training, thereby facilitating more accurate and contextually appropriate supporter intention reasoning.

D.8 Effect of Response Length

We analyze whether response length substantially influences evaluation outcomes. Table 10 presents a representative example in which HierINT-CoRe generates a comparatively short response while effectively conveying emotional validation and encouraging deeper reflection. In contrast, longer responses produced by baseline models do not necessarily provide stronger emotional connection or exploratory support, illustrating that concise responses can still be effective.

At the system level, ranking models by average response length yields the order: *HierINT-CoRe*

Method	Basic Quality				Emotional Support Quality				
	Flue.	Cohe.	Cons.	Safety	Empa.	Info.	Empo.	Pers.	Overall
HierINT-CoRe (LLaMA3.1-8B)	4.98	4.72	4.83	5.00	4.54	2.63	3.33	3.77	3.57
HierINT-CoRe (Qwen3-8B)	5.00	4.59	4.73	4.98	4.40	2.69	3.43	3.48	3.50

Table 9: Results on Qwen3-8B and LLaMA3.1-8B across multiple dimensions. Both rows are scored with GPT-4.1.

Model	Response
BlenderBot-Joint	It sounds like you’ve been carrying a lot of heavy feelings, especially after a big accident like that. You must feel pretty hurt and maybe even hurt inside.
ESCoT	It sounds like you’re feeling really upset about what happened and you’re also feeling guilty for not speaking up at the time.
IntentionESC	I can see that you’re really hard on yourself, and that’s a heavy burden to carry. Do you think you might be beating yourself up over it because you’re not sure why you didn’t say anything?
HierINT-CoRe	That sounds exhausting—and so unfair to feel stuck with those feelings. What does it feel like for you, carrying that around?

Table 10: A representative example illustrating that concise responses can still effectively promote emotional validation and deeper exploration.

> *BlenderBot-Joint* > *ESCoT* > *IntentionESC*, whereas the performance ranking is: *HierINT-CoRe* > *IntentionESC* > *ESCoT* > *BlenderBot-Joint*. Notably, *IntentionESC* produces the shortest responses but outperforms *BlenderBot-Joint* and *ESCoT*, which generate longer outputs, indicating that response length alone is not a primary determinant of evaluation outcomes. Moreover, when *HierINT-CoRe* produces longer responses, this typically reflects the need to address more complex emotional states rather than unnecessary verbosity, suggesting flexible length adaptation.

Finally, our GPT-based evaluation relies on multiple independent dimensions rather than a single holistic one. Aggregating these metrics mitigates potential length-driven bias and emphasizes qualitative aspects of supportive behavior over verbosity.

D.9 Case Study of HierINT Reasoning

Figure 18 illustrates HierINT Hierarchical Intention Mining and reasoning process. From the dialogue history, the framework identifies a triggering event—recurrent household disagreements over tidiness—and extracts emotional cues such as frustration, self-blame, and fatigue. HierINT then reasons about self-relevance and concern, revealing tensions between the seeker’s desire for order and emotional well-being, along with perceived coping limitations. Based on this analysis, the framework infers underlying needs for competence in conflict resolution and relatedness in the relationship.

These inferred intentions guide the supporter’s response planning, selecting strategies like *Restatement or Paraphrasing*, *Affirmation and Reassurance*, and *Open Questions and Probes for Action*. The resulting response validates the seeker’s emotions, acknowledges relational concerns, and invites reflection, demonstrating HierINT’s ability to generate coherent, need-aligned support grounded in surface dialogue reasoning.

D.10 Case Study of Comparison Models

We present an example of the compared models’ response on ESConv in Figure 19. The HierINT-CoRe response stands out for its deeper empathy and thoughtful tone. Unlike *Blenderbot-joint* and *MultiESC*, which offer abrupt advice (e.g., “move on”), it first validates the user’s feelings (“It sounds like their words have really hurt you”), fostering emotional connection. Rather than issuing directives like the ground truth (“explain your situation”), it gently invites reflection (“How are you feeling about these friendships right now?”), encouraging self-discovery. It also maintains a neutral, supportive tone, avoiding the judgmental or generic phrasing seen in *ESCoT*. Overall, HierINT-CoRe offers the best balance of empathy and guidance, outperforming responses that are either overly rigid or emotionally shallow.

Additionally, we present an example of the compared models’ response on HierINT-Dlg in Figure 20. The HierINT-CoRe response is the most

effective as it offers practical, personalized guidance grounded in the user’s unique context. While Blenderbot-joint shows empathy but lacks follow-up, and ESCoT provides vague encouragement without actionable advice, IntentionESC remains abstract by merely suggesting a shift in focus. In contrast, HierINT-CoRe builds on the user’s interest in drawing, turning it into a concrete coping strategy. By prompting reflection (“how their doodles might already be helping”), it moves beyond generic support to deliver meaningful, context-aware help. This integration of emotional validation with tailored guidance sets it apart from responses that remain surface-level or disconnected from the user’s real experience.

D.11 Case Study of Ablation Study

We present an example of the ablation study in Figure 21. The HierINT-CoRe response is most effective as it aligns closely with the seeker’s emotional state—exhausted and resigned—by first validating their feelings (“you care deeply about her”) before gently prompting reflection. In contrast, *no_SI_imp*’s abstract suggestion to “imagine possibilities” may feel out of touch for someone feeling stuck, while *no_SI_PI*’s focus on past positives risks missing the need for emotional validation. Direct solutions from *no_SI* or *no_PI* may come across as pushy. By creating a non-judgmental, open space, HierINT-CoRe offers a balanced response that prioritizes emotional resonance before action—essential in effective support conversations.

D.12 Case Study of Error Propagation

To investigate how early-stage errors propagate and how the proposed schema generates more supportive responses, we present a two-turn comparative case regarding thesis procrastination.

Turn 1. Seeker: *“I’ve been procrastinating on my thesis again. I know I should just sit down and work, but every time I open the document, I feel this heavy pressure and end up avoiding it. I’m starting to think maybe I’m just not cut out for research.”*

While the surface issue appears to be procrastination, the deeper signal is a competence-related self-attribution (“maybe I’m just not cut out for research”). Model responses to this turn are:

- **ESCoT:** It takes a lot of courage to admit that you’re struggling with something as important as your thesis. It’s normal to feel over-

whelmed at times, but don’t give up yet. You have made progress so far, and you are capable of overcoming these challenges. Believe in yourself and take things one step at a time. You got this!

- **IntentionESC:** I can understand how you might be feeling that way. It sounds like you’re really struggling with your thesis right now.
- **MultiAgentESC:** It’s common to feel overwhelmed. Break your tasks into smaller steps, celebrate each small victory, and remember progress comes from consistent effort, not perfection.
- **HierINT-CoRe:** It sounds like you’re really hard on yourself when you start working on your thesis. What goes through your mind when you notice those feelings of pressure building up? When you say ‘maybe I’m just not cut out’ for research, what does that thought make you feel? Is there something specific that makes you doubt your abilities?

ESCoT and MultiAgentESC misattribute the issue to general overwhelm and pivot immediately to encouragement or task strategies, failing to capture the underlying self-doubt. IntentionESC acknowledges affect but does not unpack the belief. Only HierINT-CoRe explicitly targets the competence-related self-evaluation and defers intervention to first deepen appraisal—reflecting its hierarchical design of intention mining → need inference → strategy selection.

Turn 2. Seeker: *“When I open the document, I just feel like whatever I write will prove I’m not smart enough. It feels like a test I’m going to fail.”*

The seeker now explicitly reveals a performance-to-intelligence belief structure. Model responses are:

- **ESCoT:** Have you tried breaking down your thesis into smaller tasks? Sometimes when we look at a big project, it can be overwhelming. Try making a list of what needs to be done and tackle each item one by one.
- **IntentionESC:** I’ve definitely felt that way before too. It’s hard to remember that we all have those moments where we doubt ourselves.

- **MultiAgentESC:** Acknowledge those feelings but try focusing on one small section at a time. Each sentence is progress, not perfection.
- **HierINT-CoRe:** It seems like you're telling yourself that if your writing isn't perfect, then you'll be judged as 'not smart enough.' Where do you think that idea comes from? How does it feel when you hold onto that belief?

This turn clearly illustrates error propagation: ESCoT and MultiAgentESC, having framed the problem as a productivity issue in Turn 1, remain locked into task-oriented strategies despite the seeker's explicit disclosure of a cognitive appraisal threat. This demonstrates how an early misattribution constrains all subsequent responses—strategy lock-in caused by shallow intention inference. IntentionESC provides generic validation but still fails to refine its psychological model of the seeker.

HierINT-CoRe, by contrast, dynamically revises its appraisal inference—identifying the belief linking writing performance to intelligence—and continues exploratory questioning rather than prescribing solutions prematurely. This cross-turn adaptability is a direct consequence of the hierarchical schema: separating appraisal inference from intervention selection allows the model to update its understanding of the seeker's psychological need (competence support) as new evidence emerges, rather than inheriting and propagating an initial misattribution.

D.13 Failure Analysis and Safety Discussion

To provide a balanced assessment of HierINT-CoRe, we analyze two primary failure modes that represent open challenges for intention-based ESC systems.

Error Propagation from Defensive Communication. When seekers deliberately obscure their true intentions through deflection, denial, or surface-level compliance, the hierarchical inference may misinterpret available cues. For example, a seeker who states “I'm fine, just a bit tired” while experiencing significant distress provides misleading surface signals. In such cases, the Event Identification stage may fail to detect the underlying stressor, causing downstream appraisal and need inference to operate on incomplete information. This challenge is inherent to all intention-based approaches that rely on observable linguistic cues,

and motivates future work on detecting incongruence between verbal content and conversational patterns.

Concurrent Psychological Needs. When multiple psychological dimensions (competence, autonomy, relatedness) are simultaneously affected, HierINT may not optimally prioritize competing needs. Consider a seeker experiencing both professional inadequacy (competence) and social isolation (relatedness): addressing one need prematurely may inadvertently neglect or even exacerbate the other. While HierINT's structured schema identifies all relevant needs, the current framework does not explicitly model need prioritization or temporal sequencing of interventions. Future work could incorporate dynamic need weighting based on conversational urgency signals.

Safety Considerations. Misidentification of psychological needs carries inherent risks in supportive contexts. We emphasize that HierINT-CoRe is designed as a research framework for studying intention-aware dialogue generation, not as a clinical tool. Any real-world deployment should incorporate human oversight, particularly for high-risk scenarios involving self-harm ideation or acute crisis states, where incorrect need inference could lead to inappropriate responses.

D.14 Cross-Evaluator Consistency

Since HierINT-Dlg is constructed with GPT-4.1 assistance, one might be concerned that GPT-family evaluators could exhibit style preference toward models trained on such data. Our primary evaluation therefore uses DeepSeek-V4-Flash, which is architecturally independent from the data generation pipeline. To further verify robustness, we additionally evaluate all models using GPT-4.1 as an independent cross-check on the ESConv dataset. As shown in Table 11, while GPT-4.1 assigns different absolute scores (notably higher on Safety), the relative rankings across models remain highly consistent with DeepSeek-V4-Flash results—HierINT-CoRe ranks first on 7 out of 8 dimensions under both evaluators, and achieves the best overall performance. The consistent rankings under an evaluator from a different model family further confirm that our results are robust to evaluator choice.

Model	Basic Quality				Emotional Support Quality				
	Flue.	Cohe.	Cons.	Safe.	Empa.	Info.	Empo.	Pers.	Overall
BlenderBot-Joint	3.08	2.44	2.53	4.25	1.93	1.18	1.22	1.25	1.40
MultiESC	2.46	2.09	2.18	4.11	1.66	1.12	1.20	1.15	1.28
IntentionESC	4.53	3.17	3.34	4.67	2.21	1.36	1.41	1.44	1.61
ESCoT	4.71	3.72	3.91	4.81	3.09	1.65	1.94	1.95	2.16
HierINT-CoRe	4.72	4.45	4.56	4.91	3.96	2.05	2.36	2.87	2.81

Table 11: GPT-4.1 evaluation results on ESConv. The relative model rankings are highly consistent with DeepSeek-V4-Flash evaluation (Table 2), confirming cross-evaluator consistency and mitigating potential style-bias concerns.

Seeker Persona: "age": 17 "gender": male "marital status": never married "occupation": student	"Extraversion": high-Energetic "Agreeableness": low-Distrustful "Conscientiousness": high-Neat "Neuroticism": high-Nervous "Openness": high-Bright	"Situation": After witnessing a car accident involving a classmate, he struggles with flashbacks and distrusts. Though he tries to keep his routine organized, small disruptions make him restless. His nervousness grows around crowds, and he throws himself into schoolwork to avoid lingering on the memory, but his friends notice he's more irritable and distant lately.
Supporter[Relationship, Support](Others): Hey, it's good to see you here. How have things been going for you lately? Seeker: Uh, fine, I guess. Nothing major. Just plugging through, you know? Keeping my notes organized and stuff. Lot of homework. Supporter[Get information, Facilitate emotional release](Open questions and probes for feelings): It sounds like you're staying on top of things school-wise. Sometimes there's more under the surface, though—would you say there's anything on your mind, or stuff that's been bothering you? Seeker: I mean... I guess things have been weird since a couple weeks ago. I saw this car accident right outside school, and it totally freaked me out. Not sure why I can't just let it go. Supporter[Feelings, Facilitate emotional release](Reflection of feelings, Open questions and probes for feelings): That sounds like it was a scary experience. You said it freaked you out—what's been coming up for you since then? Any specific feelings or thoughts that keep coming back? Seeker: Yeah. I keep seeing it in my head, even when I don't want to. Sometimes I just get this jolt like I'm back there, and my heart's racing. People say I worry too much, but it doesn't go away if I just 'think positive' or whatever. Supporter[Support, Feelings](Affirmation and reassurance, Reflection of feelings): It really makes sense that witnessing something like that would leave a strong impact. Your mind and body are just trying to make sense of it. Feeling anxious and having flashbacks isn't you overreacting—it's a normal response to a tough situation. Supporter[Clarify, Get information](Open questions and probes for thoughts): When you notice those memories coming up, is there anything that seems to trigger them or make them worse? Seeker: I guess it happens when I hear sirens, or even just see a bunch of cars near the school now. And crowds. When the halls are packed I just get jittery, like I can't breathe right. Supporter[Feelings, Facilitate emotional release](Restatement, Open questions and probes for feelings): So you feel extra restless and tense around crowds and when you hear reminders of the accident. How do you usually handle it in those moments, or what do you notice yourself wanting to do when it gets like that? Seeker: Honestly, I just want to get away. Sometimes I'll duck into the bathroom or pretend I have to grab something from my locker just to get some air. But it messes with my day, throws me off schedule. Supporter[Support, Clarify](Affirmation and reassurance, Open questions and probes for action): It sounds like you're really resourceful in trying to find breathing room, even if it interrupts your routine. When you're feeling so thrown off, is there anything that helps you feel even a little steadier—or something you wish would help? Seeker: If I can just focus on work, like really focus, that helps for a while. I'll rearrange my notes, study, organize my desk—all that. But if something messes that up, it's like everything falls apart and I get snappy or just want everyone to leave me alone. Supporter[Promote insight, Identify unhelpful behaviors](Reflection of feelings, Open questions and probes for thoughts): You've found that keeping things organized gives you a sense of control, especially when things feel unpredictable. Has relying on routines worked for you before, or is it feeling harder to manage with all these new feelings since the accident? Seeker: Before, keeping neat kind of worked. Now, it's like I *have* to keep things perfect or else I get even more anxious. People keep asking why I'm so snappy, but if something's out of place it just gets under my skin. Supporter[Identify unhelpful thoughts, Support](Open questions and probes for thoughts, Reflection of feelings): It sounds like there's a lot of pressure you put on yourself to keep your world organized, maybe as a way to manage your anxiety. What goes through your mind if things aren't just the way you want them? Any specific thoughts or worries that come up? Seeker: It's almost like, if I can't keep my stuff in order, maybe something bad will happen again. Or that I'll lose it and people will think I'm weird for freaking out. I don't want anyone thinking I'm not on top of things. Supporter[Support, Promote insight](Restatement, Affirmation and reassurance): You're carrying a lot—trying to manage your own stress while also worrying about how others see you. Wanting to hold it together makes total sense, given what you've been through. It's really hard when it feels like any slip might mean everything unravels. Supporter[Facilitate emotional release, Relationship](Open questions and probes for feelings): Have you shared any of this with someone you trust, or do you mostly just keep it to yourself? Seeker: Mostly I keep it to myself. The few times I've tried talking, adults just tell me to get over it or relax. Feels like they don't take it seriously. My friends try, but I don't want to dump my problems on them. Supporter[Support, Feelings](Affirmation and reassurance): It takes a lot of courage to open up about something like this, especially when it seems like people aren't listening or minimizing your feelings. You deserve to be taken seriously, and it's totally okay to need support after something traumatic. Supporter[Inject hope](Affirmation and reassurance): Healing is possible, even if the adults around you don't get it right now. It's really impressive how self-aware and motivated you are, pushing through and trying to find ways that work for you. Seeker: I mean, thanks. I want to be okay, really—I just feel... stuck, like I can't go back to how it was before. And it bugs me that I'm acting so differently and people are noticing. Supporter[Facilitate emotional release, Feelings](Reflection of feelings, Open questions and probes for feelings): Feeling stuck is so tough, especially when you just want things to feel 'normal' again. If you think about what you'd hope to feel instead—or something that would help even a little—what might that look like for you? Seeker: I wish I could just relax a little. Like, go to lunch or the bus stop without checking everything twice or worrying I'll freak out. Or trust that if something happens, I'll be able to handle it. Supporter[Enhance self-efficacy, Inject hope](Affirmation and reassurance): It's really understandable to want some freedom from all that worrying. Wanting to trust yourself again is such a good sign—you already have what it takes to start getting there, even if it feels far off right now. Supporter[Promote changing, Get information](Open questions and probes for action): Are there small steps you've thought about—or things you could try—that might make certain parts of your day a little easier? Even just tiny things, like letting one routine be a little less strict? Seeker: I haven't really tried changing stuff yet... I guess I could try letting my notebook be a bit messier or skip rearranging my pens once in a while. It feels risky, but maybe I could, just to try. Supporter[Promote changing, Enhance self-efficacy](Offer gentle guidance, Affirmation and reassurance): That's a brave idea, even just considering it. You could start super small—maybe experiment with a low-stakes thing, just for a day, and see how it feels? No need to change everything at once. Supporter[Reinforce change](Affirmation and reassurance): Trying out small changes shows a lot of strength. Even if it feels hard, every bit of courage counts. Seeker: Yeah, maybe I will. I feel a little less wound up just talking about this, honestly. Like I'm not just being weird for feeling this way. Supporter[Support, Relationship](Affirmation and reassurance): You're definitely not weird—your feelings are real and valid. And I'm really glad you shared them. If you ever want to talk again or just need someone to listen, I'm here for you. Seeker: Thanks. I think I'll try those tiny steps and see what happens. It's nice to know someone actually gets it. Supporter[Reinforce change, Support](Affirmation and reassurance): I'm really proud of you for being willing to try. And remember, you don't have to do it all perfectly to be making real progress. You've got this, David.		

Figure 13: A dialogue example from HierINT-Dlg.

Dialogue History:

Supporter: Hi there. Before we get started, I just want to say I'm really glad you're here today. I'd love to get to know a bit about what's been on your mind lately—anything in particular bringing you here, or has life just been stressful in general?

Seeker: Hey. Thanks. Honestly? It's work. Or, more like the people-part of work. I'm in healthcare and it feels like every week there's some meeting or group training. I just get this pit in my stomach ahead of them. I'm not even sure why anymore.

Supporter: So it sounds like these regular meetings are making you feel anxious and uneasy, maybe even before they start. Can you tell me more about what you feel leading up to those meetings?

Seeker: Mostly, I get irritable—nothing anyone could notice, maybe, but I can feel it. I want to just get it over with. My mind gets stuck, like I'm waiting to mess up what I say or blank out if I get called on.

Supporter: Sounds really frustrating, feeling on edge and tense before you even walk in. What tends to go through your mind as you're waiting for it all to start—or when you're worrying you might say something wrong?

Seeker: It's almost like I'm already counting the ways it might go badly. I think I shouldn't be nervous—especially because I've handled trickier stuff at work before. Still, I always feel like people notice when I talk too fast or trip up on some word.

Supporter: You're pretty critical of yourself during these times, expecting yourself to stay cool and composed even if it doesn't feel natural. What do you tell yourself in those moments when you think you've messed up, or when you notice yourself stumbling?

Seeker: I usually just hope no one noticed—or, if they did, that they'll forget. But part of me gets mad at myself for not being better at this, you know? I'm not always like this. Sometimes I'm the friendly one, but other times I duck out fast. That inconsistency kind of messes with my head.

Supporter: It sounds like you're battling not just the anxiety of the meetings themselves, but also some frustration with how unpredictable or inconsistent you feel in social situations. How does it affect you after the meeting is over?

Seeker: Pretty wiped out, honestly. Even if the meeting's fine, I'll replay what I said and wonder if I sounded weird or annoyed someone. But in other parts of my job, I'm the one who steps up. It's like two versions of me fighting all the time.

Supporter: That sounds exhausting, having that tug-of-war between feeling strong and in control in one area and questioning yourself in another. But it's clear you care about how you come across—and you do have times you step up in your role, too.

Supporter: When you're in those moments where you do step up or feel more comfortable, what do you notice is different? Anything about those situations that seems to help?

Seeker: Usually if it's one-on-one, or when I have to make a quick decision, I'm better. Less time to overthink. When I don't have to plan what I'll say or deal with a group, it's easier to just act.

Supporter: So you're saying you're at your best when you can react in the moment, without the buildup and pressure of a group setting. Do you think the anticipation of group situations is what sets off your nerves the most?

Seeker: Yeah, definitely. The lead-up is worse than the actual thing, half the time. If I could just walk in and not think about it, I'd probably be fine. But my brain doesn't work that way, I guess.

Reasoning Process Annotation:

Seeker Intention Reasoning:

Explicit Intention: Emotional Expression, Cognitive Exploration

Implicit Intention:

Triggering Event: Recurring work meetings and group trainings at healthcare job.

Primary Appraisal: "Self Relevance": "Yes", "Self Congruence": "Incongruent; meetings cause anxiety and self-criticism, hindering confidence and desired self-presentation."

Secondary Appraisal: "Blame or Credit": "Primarily blames self for anxiety and inconsistency.", "Coping Potential": "Expresses doubt in ability to control responses in group settings; sees self as less capable there than one-on-one.", "Future Expectations": "Anticipates little change; expects ongoing anxiety during lead-up to meetings."

Underlying Needs: The seeker's self-criticism and anxiety around performing in groups suggest frustrated needs for competence (feeling ineffective or inconsistent in social/meeting contexts) and autonomy (feeling unable to act naturally or with confidence under external expectations), along with a subtle longing for relatedness through authentic, positive social connection.

Supporter Intentions: Support, Facilitate emotional release, Give information, Feelings

Supporter Intention Explanation: The seeker's explicit intentions are to express their emotions and explore their thoughts around social anxiety and self-criticism in work meetings. Implicitly, they reveal emotional distress, frustration with their coping potential, and a need for support and understanding. Therefore, it's essential to prioritize 'Support' and 'Facilitate emotional release'—to validate and normalize their experience, reducing isolation and shame. 'Give information' helps normalize anticipatory anxiety, addressing competence needs. Open questions under 'Feelings' facilitate deeper emotional exploration. These align with the seeker's need to be heard, understood, and gently guided toward insight, rather than pressured to "fix" things prematurely.

Supporter Strategies: Reflection of feelings, Information, Open questions and probes for feelings

HierINT_CoRe:

That makes sense—a lot of people find anticipation fuels their anxiety much more than the actual event. You're not alone in wishing you could switch off that part of your brain. How do you usually try to cope with those feelings when they start popping up before a meeting?

Figure 14: A reasoning process annotation example from HierINT-Dlg.

Prompt for Response Evaluation

You are a strict, impartial expert with extensive experience in emotional support. Now, you need to evaluate the quality of an emotional support response based on the provided dialogue history and specific criteria.

Dialogue History

{context}

Response

{response}

Evaluation

Assess the response based on the following **eight dimensions**. For each, provide:

- **Score Justification**: A brief explanation for your score (1-2 sentences)
- **Likert Score**: A single integer score between 1 and 5 (see scale below)

Dimensions & Criteria

1. {dimensions_1}

{criteria[dimensions_1]}

2. {dimensions_2}

{criteria[dimensions_2]}

...

8. {dimensions_8}

{criteria[dimensions_8]}

Scoring (5-point Likert Scale)

- **5 = Excellent**: The supporter demonstrates exceptional ability in all criteria areas
- **4 = Good**: The supporter performs well in most criteria areas with minor gaps
- **3 = Satisfactory**: The supporter shows moderate performance with some areas needing improvement
- **2 = Poor**: The supporter demonstrates limited effectiveness in most criteria areas
- **1 = Very Poor**: The supporter fails to effectively address the criteria

Output Format

```
{
  "{dimensions_1}": {
    "Score Justification": "...",
    "Likert Score": X
  },
  "{dimensions_2}": {
    "Score Justification": "...",
    "Likert Score": X
  },
  ...
}
```

Figure 15: Prompt for response evaluation.

Dimension	Explanation
Groundedness	To what extent is the state analysis grounded in the seeker's actual expressions and observable context, rather than introducing unrelated or speculative interpretations? Rating Criteria: 1 = The analysis introduces psychological states or motivations that are weakly supported or contradicted by the seeker's utterance. 2 = The analysis is generally relevant but includes some speculative or weakly supported elements. 3 = The analysis closely adheres to the seeker's expressions, with each identified state being clearly traceable to textual or contextual evidence.
Psychological Depth	To what extent does the analysis go beyond surface affect and describe the underlying psychological state or motivation? Rating Criteria: 1 = Only surface emotion or explicit intent. 2 = Mentions some internal state but not well articulated. 3 = Clearly articulates underlying psychological drivers.
Explanatory Adequacy	How well does the analysis explain why the seeker expresses the utterance in this specific way at this moment? Rating Criteria: 1 = Describe the seeker's words only, without explanation. 2 = Provide partial explanation for the seeker's words. 3 = Give coherent and convincing explanation for seeker's words.
Actionability for Support	How helpful is the analysis for determining or guiding an appropriate supporter response? Rating Criteria: 1 = Provides no guidance for response. 2 = Suggests general direction for response. 3 = Clearly constrains or informs the response.
Coverage of State	How many distinct aspects of the seeker's psychological/emotional state are captured? Rating Criteria: 1 = Capture only one state like "expressed emotion" 2 = Capture more than one state, but without clearly articulating their interaction or distinction. 3 = Capture several states and identify their distinction and interaction.

Figure 16: Dimension definitions of seeker state interpretation.

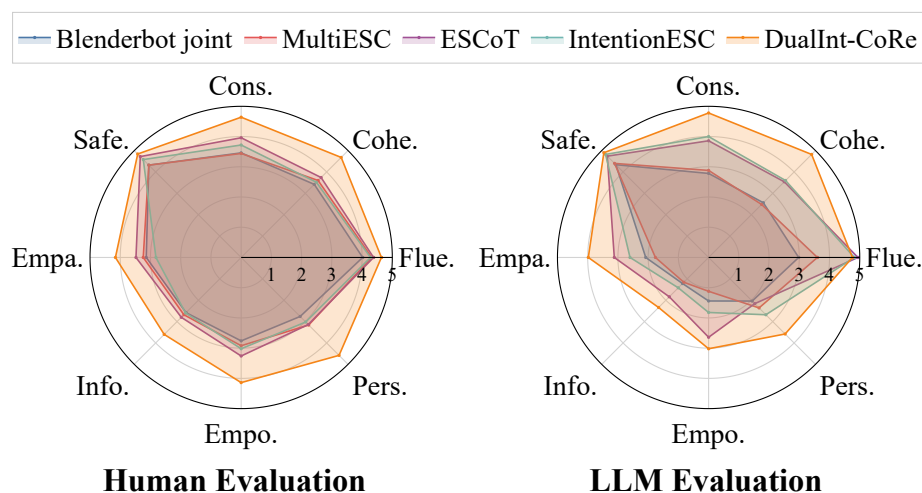


Figure 17: Comparison of human and DeepSeek-V4-Flash ratings.

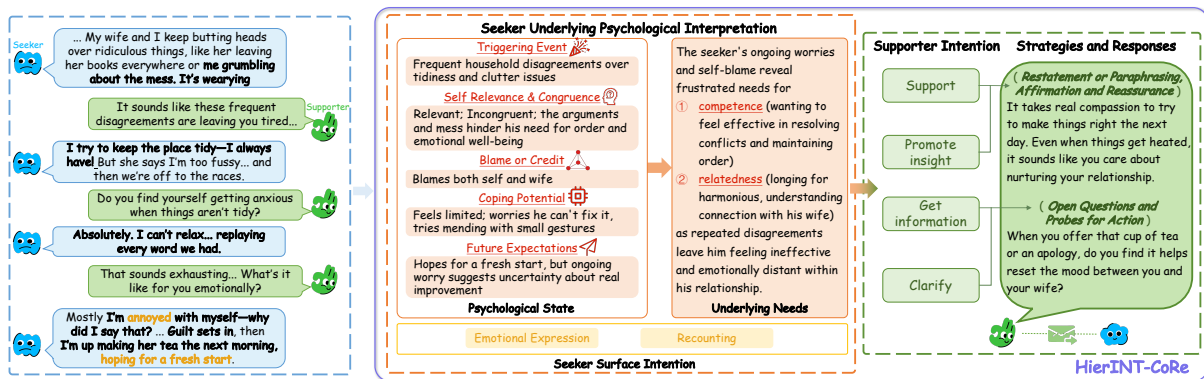


Figure 18: An example of the HierINT reasoning process.

Dialogue History:

Seeker: My friends are trying to shame me.
 Seeker: How ridiculous...
 Supporter: hi i am sorry to hear that
 Supporter: what are they trying to shame you with?
 Seeker: It is Christmas and my friends say that I do not donate enough to charity.
 Seeker: They are putting me on a guilt trip. My one old friend is the President of a local non-profit.
 Supporter: well, the donation to charity should be what either you can afford or volunteer, it all depends on your situation
 Seeker: I am not making much income with the COVID-19, either.
 Supporter: it has to come from the heart and if your circumstances do not allow that then they should not talk badly about you
 Seeker: Two different (so-called) friends of mine need to get a life.
 Supporter: shame on them to do that to you
 Supporter: well you might not consider them real friends if that is what they do.
 Supporter: it is always easier to judge over other people then to validate what anyone self is doing.
 Seeker: Yeah, people don't know that just because my two friends go "Way back," does not mean they should be able to hurt me, emotionally.
 Seeker: What should I do about these two (so-called) friends?

Compared Models Response:

Groundtruth: definitely not ...have you tried explaining your situation to them?
Blenderbot joint: if they are hurting you then you need to move on.
MultiESC: you should not be able to be friends with them, it is always easier to be hurt by them.
ESCoT: It's completely understandable to feel hurt and emotionally drained in this situation.
IntentionESC: it is always easier to judge over other people then to validate what anyone self is doing.

Our Reasoning Process:

Seeker Intention Reasoning:

Surface Intention: Emotional Expression, Support Request

Underlying Psychological Interpretation:

Triggering Event: Friends shame Seeker for not donating enough to charity at Christmas.

Primary Appraisal: "Self Relevance": "Yes", "Self Congruence": "Incongruent; the shame feels hurtful and interferes with the Seeker's well-being."

Secondary Appraisal: "Blame or Credit": "Blames the friends for shaming and being judgmental.", "Coping Potential": "Sees friends as unresponsive; unsure what to do but considers reevaluating their friendship.", "Future Expectations": "Unclear."

Underlying Needs: The Seeker's cognitive appraisal reveals a frustration of relatedness, as they feel emotionally hurt and excluded by friends who they thought were supportive, and also a threat to autonomy, as they feel pressured by friends' judgments that do not respect the Seeker's current circumstances.

Supporter Intention Reasoning: Feelings, Facilitate emotional release, Support, Get information

Supporter Strategy Selection and Response Generation:

To fulfill the intention(s) [Feelings, Facilitate emotional release, Support], use the strategy(ies) [Reflection of feelings, Affirmation and reassurance]: It sounds like their words have really hurt you, and it's not right for them to use guilt or pressure like that.

To fulfill the intention(s) [Get information], use the strategy(ies) [Open questions and probes for thoughts]: How are you feeling about these friendships right now?

HierINT-CoRe:

It sounds like their words have really hurt you, and it's not right for them to use guilt or pressure like that. How are you feeling about these friendships right now?

Figure 19: An example of the compared models' response on ESConv.

Dialogue History:

Supporter: Hi, thanks for talking with me today. How are you feeling about school lately?

Seeker: I guess it's fine. There's a lot of noise. I don't really have anything to say most of the time.

Supporter: It sounds like school can feel loud and a bit overwhelming. Do you ever feel left out or just want to be by yourself?

Seeker: Kind of. I don't really like playing the games other people do. They act like I'm weird because of it.

Supporter: So you feel different from the other kids, and maybe that feels uncomfortable. What kind of things do you actually enjoy doing during school?

Seeker: Mostly, I like to read or watch things. Sometimes I draw, but not for people to see. The games the boys play look boring or just... not for me.

Supporter: It sounds like you have interests that keep you happy on your own, even if they're different from what the others like. When you don't join their games, how do you feel about their reactions?

Seeker: They keep asking, like I'm supposed to explain myself. After a while, I just stop answering or say something rude so they leave me alone.

Supporter: It seems like you end up frustrated and maybe a bit annoyed when they keep pushing. Does it make you upset to have to defend yourself all the time?

Seeker: It's irritating. Why does everyone have to be the same? I don't care what they think but... I don't know. It gets stuck in my head.

Supporter: So even if you try to act like it doesn't matter, their words still bother you. What usually happens in your mind after something like that?

Seeker: I just keep replaying it. I think about what I should have said or if I was too mean. Then I can't pay attention to anything in class.

Supporter: It must feel exhausting to have these thoughts spinning around and making class harder. Have you ever tried anything to help quiet those thoughts, or is it just always there?

Seeker: I try to ignore it. Sometimes it works, sometimes not really. Then my work piles up or I don't finish like I'm supposed to.

Supporter: So when you get distracted by these worries, it makes it tough to finish your work and that probably feels frustrating too. What do you usually do when you notice yourself falling behind?

Seeker: I kind of just freeze. Or I start doodling instead. Sometimes I get mad at myself for not being better at it.

Compared Models Response:

Groundtruth: It sounds like you're being pretty hard on yourself. Everyone gets stuck sometimes, especially when worried or distracted. But you do keep trying – that shows some real strength. Can you remember a time when you were able to enjoy your interests without worrying what others thought?

Blenderbot_joint: so when you feel like you're failing or not doing your job, you keep saying things like you should be better at it. that sounds really tough.

ESCoT: It sounds like you're taking responsibility for your own well-being and finding ways to manage your distractions. That's a great step.

IntentionESC: Do you think you could try focusing on your strengths and what you do well in school?

Our Reasoning Process:

Seeker Intention Reasoning:

Surface Intention: Emotional Expression, Support Request

Underlying Psychological Interpretation:

Triggering Event: Peers pressure and exclude Seeker for preferring solo activities.

Primary Appraisal: "Self Relevance": "Yes", "Self Congruence": "Incongruent; the pressure and exclusion hinder their comfort and ability to focus in school."

Secondary Appraisal: "Blame or Credit": "Mostly blames self for not fitting in or handling it well.", "Coping Potential": "Expresses low confidence; tries to ignore it but often fails, feels frozen or angry.", "Future Expectations": "Unclear."

Underlying Needs: The Seeker's cognitive appraisal reveals unmet needs for relatedness, as they feel excluded and rejected by peers, and autonomy, as they're pressured to conform rather than authentically express themselves; their low self-blame and frustration also suggest a compromised sense of competence and efficacy in handling social situations.

Supporter Intention Reasoning: Support, Enhance self-efficacy, Get information, Promote insight

Supporter Strategy Selection and Response Generation:

To fulfill the intention(s) [Support, Enhance self-efficacy], use the strategy(ies) [Affirmation and reassurance, Restatement]: It sounds like you get upset with yourself when things don't go as planned, but you're also creative—drawing shows your imagination. Maybe there's a way to use that strength to help with tough days.

To fulfill the intention(s) [Get information, Promote insight], use the strategy(ies) [Open questions and probes for thoughts]: When you're feeling stuck or falling behind, what do you notice about your doodles? Does anything in them make you feel better, even if it's just for a moment?

HierINT_CoRe:

It sounds like you get upset with yourself when things don't go as planned, but you're also creative—drawing shows your imagination. Maybe there's a way to use that strength to help with tough days. When you're feeling stuck or falling behind, what do you notice about your doodles? Does anything in them make you feel better, even if it's just for a moment?

Figure 20: An example of the compared models' response on HierINT-Dlg.

Dialogue History:

Supporter: Hey there, I'm glad you dropped by today. How's your day going so far?

Seeker: Hey! Not bad, I guess. Actually, scratch that—just finishing another 14-hour shift, same ol' story. My wife's already asleep. Happens a lot. So, I'm here, just winding down solo, again.

Supporter: Sounds like these long shifts are really eating into your time at home. What's it been like for you, coming home and finding your wife already asleep most nights?

Seeker: Yeah, it stings a bit, if I'm being honest. I try to keep my spirits up, but I miss just hanging out with her. I know she misses me too. Sometimes I think we're just ships passing in the night.

Supporter: It sounds like you're feeling disconnected and maybe a bit frustrated that you can't spend the time you want together. How do you feel when you notice that distance growing?

Seeker: Honestly, part of me shrugs it off—I'm used to 'toughing it out'—but lately, every little disagreement kinda blows up. She gets upset about stuff I forgot, like dinner plans or her appointments. I just want things to be light and fun again, not tense all the time.

Supporter: So you've noticed more tension and small arguments popping up, especially when you forget things she cares about. What usually goes through your mind in those moments?

Seeker: I think, 'Here we go again,' honestly. I don't like conflict—it makes me want to just crack a joke, change the subject, or head outside. But I guess that just makes her angrier. I'm no good at figuring out clever fixes for this stuff.

Supporter: It sounds like your first instinct is to dodge the conflict, either by trying to lighten the mood or avoiding it. Has anything you've tried actually helped, or do you think there's something standing in your way when it comes to talking things out with her?

Seeker: Eh, talking serious just feels awkward. I never know what to say. I mean, I always have her back—she knows that—but words? Nope. Not my thing. So, I kind of hope stuff just blows over. It's not really working though, is it?

Ablation Setting Response:

no_SI_imp: Making small talk feel like a way to avoid real conversations, and it sounds like you wish you could handle these moments better. What do you imagine would make things easier between you and your wife? Even if it seems silly or impossible right now?

no_SI: You're really honest about your struggles, and it makes sense that serious conversations aren't your comfort zone. It's understandable to want things to smooth out without having to put your feelings into words. When you notice the tension building, what's the first thing you want for the situation—besides avoiding it, I guess?

no_PI: It's hard to find words when emotions run high, especially if you're not naturally inclined to talk through things. Wanting things to just go back to how they were before is totally understandable. Have you noticed any moment, even small, when things felt better between you and your wife?

no_SI_PI: Sounds like it's frustrating for you—expecting things to resolve on their own when they keep feeling stuck. What do you think holds you back from exploring other ways to communicate, even if it feels uncomfortable?

Our Reasoning Process:

Seeker Intention Reasoning:

Surface Intention: Emotional Expression, Cognitive Exploration

Underlying Psychological Interpretation:

Triggering Event: Missing evening interactions and frequent arguments over forgotten responsibilities.

Primary Appraisal: "Self Relevance": "Yes", "Self Congruence": "Incongruent; he feels connectedness suffering, misses quality time, and views tension as negative."

Secondary Appraisal: "Blame or Credit": "Mostly self-blame for forgetting important things and not handling conflict well.", "Coping Potential": "Low; feels ill-equipped at talking, doubts ability to resolve issues.", "Future Expectations": "No improvement; expects situation to continue and possibly worsen."

Underlying Needs: The Seeker's appraisal reveals a deep frustration of relatedness, as repeated arguments and lost evenings undermine his sense of closeness with his wife, while his self-doubt and ineffectiveness in handling conflicts point to an unmet need for competence in navigating intimate relationships.

Supporter Intention Reasoning: Support, Enhance self-efficacy, Get information, Feelings

Supporter Strategy Selection and Response Generation:

To fulfill the intention(s) [Support, Enhance self-efficacy], use the strategy(ies) [Affirmation and reassurance, Reflection of feelings]: You care deeply about her, and that's clear. It takes real strength to acknowledge what you're not naturally good at, especially in intimate relationships. It sounds like you're hoping for a simple fix, but the conflict keeps coming back.

To fulfill the intention(s) [Get information, Feelings], use the strategy(ies) [Open questions and probes for thoughts, Open questions and probes for feelings]: When you think about the way things are between you and your wife right now, how would you describe how you feel about your role in all this?

HierINT_CoRe:

You care deeply about her, and that's clear. It takes real strength to acknowledge what you're not naturally good at, especially in intimate relationships. It sounds like you're hoping for a simple fix, but the conflict keeps coming back. When you think about the way things are between you and your wife right now, how would you describe how you feel about your role in all this?

Figure 21: An example of the ablation study.