ESC-Judge: A Framework for Comparing Emotional Support Conversational Agents

Navid Madani and **Rohini K. Srihari**
Computer Science and Engineering - University at Buffalo
Buffalo, NY, 14260
`{smadani, rohini}@buffalo.edu`

Abstract

Large Language Models (LLMs) increasingly power mental-health chatbots, yet the field still lacks a *scalable*, *theory-grounded* way to decide *which* model is more effective to deploy. We present **ESC-Judge**, the first end-to-end evaluation framework that (i) *grounds* head-to-head comparison of Emotional-Support LLMs (ES-LLMs) in an established psychological theory—Clara Hill's *Exploration–Insight–Action* (E-I-A) counselling model—thereby delivering a structured, interpretable lens on performance, and (ii) fully *automates* the pipeline at scale. ESC-Judge proceeds in three stages: (1) it synthesizes realistic help-seeker roles by sampling empirically salient attributes (stressors, personality, life history); (2) it has two candidate ES-Agents conduct separate sessions with the *same* role, isolating model-specific strategies; and (3) it asks a specialised judge LLM to issue pairwise preferences across rubric-anchored skills that exhaustively cover the E-I-A spectrum. In our empirical study, ESC-Judge matches PhD-level annotators in **85%** of Exploration, **83%** of Insight, and **86%** of Action decisions, demonstrating human-level reliability at a fraction of the cost. We release all code, prompts, synthetic roles, transcripts, and judgment scripts to catalyze transparent progress in emotionally supportive AI [1].

1 Introduction

Large Language Models (LLMs) have begun powering mental-health chatbots and peer-support apps (Stade et al., 2024). Because these agents interact with vulnerable users in high-stakes settings, the community urgently requires *rigorous, theory-grounded evaluation* to decide which models are safe and effective to deploy. Most work still probes emotional-support quality with (i) **reference-based metrics** that score responses against a single gold transcript using lexical or semantic similarity measures such as BLEU, ROUGE or BERTScore and (ii) human annotation (Zhao et al., 2023; Liu et al., 2021; Zheng et al., 2023b). Reference metrics demand large, professionally annotated corpora—expensive to create and culturally narrow—and implicitly assume one "correct" reply, ignoring the multiplicity of valid counselling strategies. Similarity metrics reward paraphrase overlap while overlooking relational depth, empathic timing, and process adherence. Finally, today's best leaderboards still lean on **live human raters**; their judgements are slow, costly, subjective and often lack expert counselling knowledge, resulting in low inter-rater agreement and poor reproducibility.

Clara Hill's *Exploration–Insight–Action (E-I-A)* framework offers an empirically validated lens on what *ought* to happen in supportive dialogues (Hill, 2014). Yet, existing benchmarks neither operationalize this theory nor test models across the diverse personalities that modulate real conversations. Moreover, their reliance on continuous expert annotation prevents scaling beyond a few hundred pairs. A truly useful benchmark must therefore (i) understand counseling theory, (ii) generalize to *many* help-seeker personas, and (iii) scale to *many* comparisons by being automated and not needing human intervention. This can enable scalable and self-supervised optimization of such agents in the future.

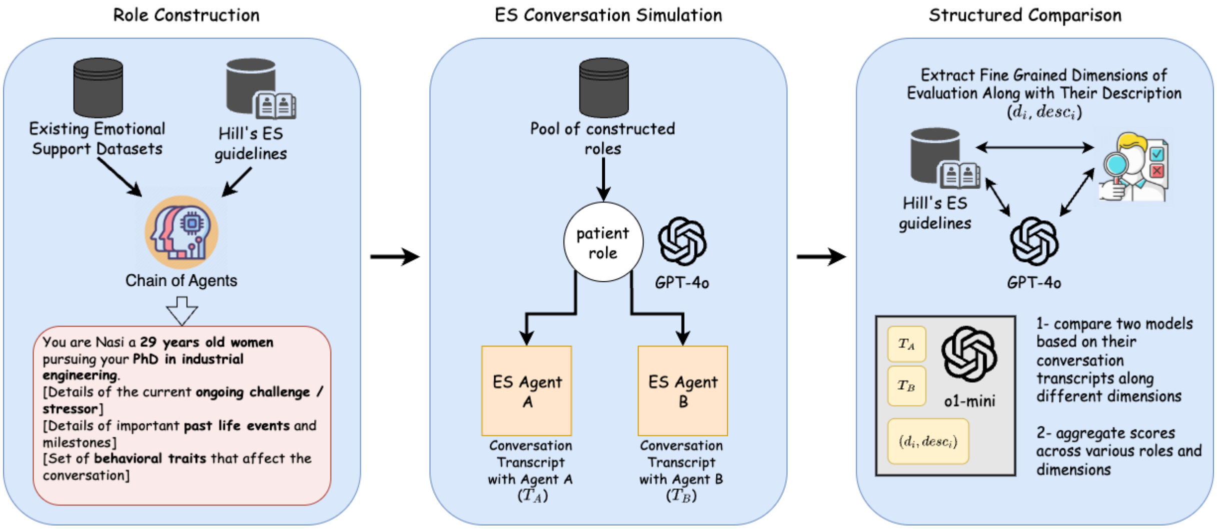
We introduce **ESC-Judge**, a three-stage, fully LLM-driven framework that addresses these gaps:

1. **Help-seeker role construction**: We sample empirically influential traits (Big Five personality, coping style, trust level, social support, triggers)—all drawn from Hill's text—to generate a spectrum of realistic help-seeker roles.
2. **Emotional support conversation simulation**: Two candidate ES models converse independently with the *same* help seeker role

---

[1] https://github.com/navidmdn/ESC-Judge

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16060–16077
November 4-9, 2025 ©2025 Association for Computational Linguistics


Role Construction
Existing Emotional Support Datasets
Hill's ES guidelines
Chain of Agents
ES Conversation Simulation
Pool of constructed roles
patient role
GPT-4o
ES Agent A
ES Agent B
Conversation Transcript with Agent A (T_A)
Conversation Transcript with Agent B (T_B)
Structured Comparison
Extract Fine Grained Dimensions of Evaluation Along with Their Description ($d_i, desc_i$)
Hill's ES guidelines
GPT-4o
T_A
T_B
o1-mini
($d_i, desc_i$)
1- compare two models based on their conversation transcripts along different dimensions
2- aggregate scores across various roles and dimensions

You are Nasi a **29 years old women** pursuing your **PhD in industrial engineering**.
[Details of the current **ongoing challenge / stressor**]
[Details of important **past life events** and milestones]
[Set of **behavioral traits** that affect the conversation]

Figure 1: Overall pipeline of our proposed ESC-Judge framework. Stage 1: constructs a diverse set of roles with various life backgrounds, demographics and behavioral attributes. Stage 2: conditioning on a fixed help seeker role, we have two emotional support (ES) models under test to participate in an emotional support conversation and we store the conversation transcripts. Stage 3: given carefully curated evaluation dimensions based on Hill's framework, we compare the capabilities of the two models under test on performing *Exploration, Insight and Action.*

under identical generation settings, isolating model-specific strategies.

3. **LLM Pairwise Judge**: A specialist judge model, instructed with the *(E-I-A)* rubric, issues `A vs. B vs. tie` preferences for each fine-grained dimension. Pairwise comparison is cognitively easier than absolute scoring, avoids ad-hoc calibration, and aligns with real-world deployment choices.

Our contributions can be summerized as follows:

- **Theory-aligned benchmark**: First end-to-end emotional support judge pipeline grounded explicitly in Hill's E-I-A counselling framework.
- **Trait-driven realism**: Introduce personality-sensitive simulation that stress-tests ES agents across diverse user profiles.
- **Scalable, expert-encoded judging**: Pairwise LLM judge achieves human-level reliability with match rate of **0.86, 0.85 and 0.83 on three categories of Exploration, Insight and Action**) while eliminating ongoing expert annotation costs.
- **Open resources**: We release code, prompts, simulated roles, transcripts, and judgment scripts to catalyze transparent progress in supportive AI.

Road-map. Section 2 reviews prior evaluation efforts; Section 3.1 details ESC-Judge; Section 4 reports experiments; and Section 5 discusses limitations and future work.

2 Related Work

2.1 Human Evaluation of Emotional-Support Dialogues

Early studies on emotional-support conversation (ESC) agents relied primarily on **human judgments**. The ESConv corpus (Liu et al., 2021) introduced theory-informed annotations of support strategies and evaluated systems via similarity measures between model utterances and human gold responses. Follow-up work continued to enlist either lay annotators or counseling experts to score generated dialogues for empathy, helpfulness, and coherence (Zhao et al., 2024; Zheng et al., 2022). Although human evaluation captures nuanced relational qualities, it is expensive, yields only moderate inter-rater agreement for subjective traits, and scales poorly to the rapid iteration cycles of modern LLMs.

2.2 Automated and LLM-Based Evaluation Protocols

Given the limitations of manual annotation, researchers have explored **automatic metrics**. Standard lexical-overlap scores (BLEU, ROUGE, BERTScore) are the most common theme (Liu et al., 2021; Zhao et al., 2023; Zheng et al., 2023b). Domain-specific proxies such as *strategy following accuracy*—predicting whether a model follows the chosen strategy correctly—offer an alternative yet important aspect which is studied in (Madani et al., 2024).

Recent advances turn large language models into *reference-free judges*. Generic dialogue benchmarks like (Dubois et al., 2024) and (Zheng et al., 2023a) prompt GPT-4 to conduct pairwise response comparisons and report moderate–high agreement with human preferences.

For the emotional-support domain, (Zhao et al., 2024) combines role-played distressed users with multi-criteria human annotation and additionally trains a ranking model (ESC-Rank) to approximate expert scores. Despite using simulated roles, the constructed roles lack nuances that affect ES conversation and the ESC-Rank model is only trained on five utterances which is significantly short for assessing the full life-cycle of an emotional support conversation.

3 ESC-Judge

3.1 Framework Overview

ESC-Judge unfolds in three sequential stages that mirror a real-world counseling encounter while enforcing strict experimental control over the patient role, the evaluation rubric and the counselor characteristics.

Stage 1 – Patient Role Construction. We construct a synthetic *help-seeker* role by sampling a bundle of empirically salient client traits, e.g., ongoing stressors, important life events, big five personality traits, etc. The resulting role prompt is injected into an instruction message that also defines session goals, ensuring every candidate model faces an identical, richly specified user role.

Stage 2 – Simulating Emotional Support Conversation Each target Emotional-Support Agent $\mathcal{M}_A$ and $\mathcal{M}_B$ engages the simulated patient in an independent dialogue session of dynamic length. This design yields two parallel transcripts whose differences stem solely from the support strategies of the competing models.

Stage 3 – LLM Judge Assessment. A specialized judge LLM receives the paired transcripts (T_A, T_B) along with an evaluation dimension and outputs a preference—$T_A \succ T_B$, $T_B \succ T_A$, or TIE when neither response is clearly superior—along with rubric-anchored rationales that score each conversation across 9 fine-grained dimensions that represent Hill's macro-dimensions (*Exploration*, *Insight*, *Action*). Together, these three tightly coupled stages deliver a reproducible, plug-and-play testbed for head-to-head comparison of ES agents while remaining faithful to Clara Hill's theoretical framework. Figure 1 demonstrates the full pipeline of our proposed framework.

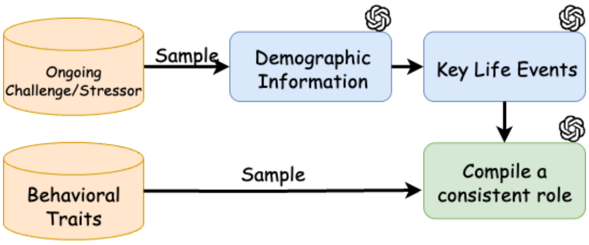


Figure 2: Role construction agents: Orange agents are random samplers based on pre-defined categories. Blue agents use generative prompts to explore the desired domain. The green agent only validates and compiles the final role without adding new information. Arrows represent the flow of data between agents.

3.2 Patient Role Construction

While designing the role-construction pipeline, we explicitly followed the factors that Hill identifies as most influential in an emotional-support encounter (Hill, 2014). This stage implements a *synthetic-role generation strategy* realized as a multi-step CHAIN-OF-AGENTS: each agent, according to figure 2 interacts with others to add a partial facet of information to enrich the proposed role. The cascade halts with a consistency agent that reviews the full role to avoid inconsistent details. We explain each of these agents as follows.

Ongoing Challenge and Stressor. Because the presenting problem anchors the entire dialogue, we begin by sampling a salient life challenge from a curated pool, collated from existing emotional support and counseling datasets (Liu et al., 2021, 2023). 6 Categories and 50 sub-categories are listed in Table 1. Given a randomly chosen category c, we uniformly sample a sub-category $s \in c$ and send it to the next agent.

Demographic Information. This agent injects essential demographic descriptors to ground the role in a credible life context. We employ a generator prompt (Chen et al., 2024) to yield a diverse yet consistent set of demographic attributes. Specifically, the agent adds gender, age, familial status, and occupation—all cross-checked for coherence with the sampled stressor. For instance, when the stressor is *Divorce or breakup*, familial status must reflect a dissolved partnership, whereas a *retired veteran* profile is never paired with an eighteen-year-old. The result is a persona whose demographic identity harmonises with the ongoing challenge, promoting realistic downstream interactions. Details about the generator prompt and the configurations used, can be found in appendix A.1.

Key Life Events. This agent imagines the help seeker's personal history by generating a ranked list of N salient life events spanning categories such as childhood trauma or positive experiences, family dynamics, romantic relationships, career milestones or failures, and loss or bereavement. Leveraging a nested generator prompt (Chen et al., 2024), it first explores a diverse set of candidate categories and then explores different scenarios within each category. It would then uniformly sample a subset to attach to the role. Prompts used for this section can be found in appendix A.2.

Behavioral Traits According to Hill's Helping Skills framework (Hill, 2014), a help-seeker's behavioral profile can profoundly shape the course and effectiveness of an emotional-support dialogue. Guided by the characteristics catalogued in the text, we organized the traits into 5 salient categories of 1) Big five personality traits 2) cognitive biases, thinking patterns and emotional baseline 3) response style towards therapist and trust in the process 4) social support network and coping mechanism and 5) triggers, sensitivities and self-soothing mechanisms each including some sub-categories. Afterwards, we sampled one representative variant from each to construct a concrete role for simulation. Table 6 lists the categories, dimensions, and exemplar variants used in our work. The selected variants along with a description are used to construct the role. More details can be found in appendix A.3.

Finally, a role construction agent takes the generated persona with demographics, key life events and sampled behavioral traits to construct a consistent full role. We utilize GPT-4o and langchain to construct the pipeline. Note that we have three types of agents in the pipeline as shown in Figure 2. Some agents only sample from a predefined data. Some are synthetic data generators and one is doing consistency check and re-writing. You can find details of each component, sample roles and prompts used for each agent in appendix A.

3.3 Simulate Emotional Support Conversation

After constructing a diverse pool of patient roles (3.2), we stage controlled dialogues to evaluate each emotional–support (ES) model under identical conditions. For every role r we create two conversations—(r, ES_A) and (r, ES_B)—so that subsequent judgments compare model behaviour given the *same* patient context. Dialogues are later scored against Hill's exploration–insight–action guidelines.

Dialogue engine The patient is realised as an autoregressive *help seeker agent*: an LLM prompted with the role card plus the running history. The support agent is the ES model under test. Agents alternate turns with fixed generation settings (temperature 0.7, top–p 0.9, max 512 tokens to generate at each utterance) to isolate model–level differences.

Turn budget and early stopping LLM pairs often spiral into repetitious closing formalities (*e.g.*, reciprocal thanks and farewells). To retain only the informative portion of the conversation, we (i) cap sessions at $T_{\max} = 20$ turns and (ii) include a lightweight logistic regression end–of–conversation detector trained on 1K dialogues based on lexical 2-gram and 3-gram utterance features. When the model classifies the utterance as end–of–conversation, we stop the conversation at that point. More details about the training and evaluation of the end–of–conversation detector can be found in appendix B.1.

3.4 LLM Judge Assessment

Interactive rubric construction. To transform Clara Hill's three macro-chapters—*Exploration*, *Insight*, and *Action*—into an operational scoring guide, we adopt a mixed LLM–human loop:

1. **Chapter parsing.** We transform the book into markdown format and clean-up the resulting text. We use (GPT_4O) and ingest each chapter and ask it to propose a candidate rubric: *a proposed dimension, its definition and behavioural anchors*.

2. **Author vetting.** Two authors independently screen the draft for faithfulness and specificity, merging identical dimensions and flagging vague ones (e.g., the initial "*exploration of feelings and thoughts*" was judged too broad. We split it into *Encouragement of Emotional Expression* and *Exploration of Thoughts and Narratives*).

Category	Sub-categories
Personal Loss & Major Life Changes	Death of a loved one; Divorce or breakup; Family estrangement; Major illness or injury; Becoming a new parent; Caring for an aging family member; Pregnancy complications; Infertility or miscarriage; Social isolation; Immigration away from family
Identity, Discrimination & Social Challenges	Exploring LGBTQ+ identity; Lack of acceptance; Racial or gender discrimination; Workplace harassment; Identity crisis; Reputation damage
Career & Academic Pressures	Job loss; Toxic work environment; Career uncertainty; Burnout; Missed promotion; Academic failure; Completing a PhD; Job relocation; Fear of automation
Financial & Economic Stress	Significant debt; Inability to pay rent; Eviction; Medical bills; Loss of savings; Living paycheck-to-paycheck; Supporting dependents; Legal financial burdens; Bankruptcy
Health & Well-being	Chronic illness; Mental-health struggles; Sleep deprivation; Major surgery; Past trauma; Eating disorders; Addiction; Medication side-effects; Terminal illness
Environmental & Societal Stressors	Moving to a new country; Natural disasters; Political unrest or war; Victim of crime; Legal trouble; Forced lifestyle change (e.g., military service)

Table 1: Stressors categories and sub-categories used during *Ongoing Challenge and Stressor* sampling.

3. **LLM clarification rounds.** For every flagged item we prompt the model with the objection and request a sharper rewrite or removal. The loop typically converges in $\leq$ 3 rounds per chapter.

4. **Pilot rating.** Annotators rate 50 dialogue pairs across all proposed dimensions with the provisional rubric; any item with agreement $\kappa < 0.5$ is re-phrased or discarded.

The final rubric contains **9 fine-grained dimensions**: *Exploration*: Encouragement of Emotional Expression, Exploration of Thoughts and Narratives, Empathic Understanding *Insight*: Establish a Trusting Foundation, Assess Readiness for Insight, Use Gentle Challenges and Interpretations. *Action*: Clarify the Desired Change, Ensure Readiness and Collaboration, Brainstorm and Evaluate Options.

Table 2 categorizes all of these dimensions along with the definition of them.

Pairwise judgement protocol. For each dimension d we feed an *o1-mini* judge with:

- The full transcripts $(T_A,\, T_B)$
- The plain-language definition of d
- A system instruction to (i) reason before judge and (ii) output a verdict: `A|B|tie`

We sample the judge twice with fixed temperature of 1.0 alternating the position of T_A and T_B to avoid position bias as emphasized by prior work (Zheng et al., 2023a). If the verdicts change in these two sets of conditions, we choose `tie` as the final verdict. If the output format does not match the prompted template (either the template is violated or the verdict is not given) we skip that instance. An example judge response is shown in appendix D.

Aggregation of Judgements. Let $w_d^{(A \succ B)} \in \{1, 0, \frac{1}{2}\}$ denote the outcome of model A versus B on dimension d

$$w_d^{(A \succ B)} = \begin{cases} 1 & \text{judge prefers A} \\ 0 & \text{judge prefers B} \\ \frac{1}{2} & \text{tie} \end{cases}.$$

Category–level comparison. For each Hill macro–category $c \in \{\textsc{Expl}, \textsc{Ins}, \textsc{Act}\}$ with dimension set D_c, the category score of A against B on a single *role* r is

$$S_c^{(A \succ B)}(r) = \frac{1}{|D_c|} \sum_{d \in D_c} w_d^{(A \succ B)}(r).$$

Across roles. Given a pool of simulated roles $\mathcal{R}$, we average:

$$\bar{S}_c^{(A \succ B)} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} S_c^{(A \succ B)}(r).$$

Decision rule. Model A is judged *preferred* to B in category c if

$$\bar{S}_c^{(A \succ B)} > \tfrac{1}{2} \quad \Longrightarrow \quad A \succ_c B,$$

otherwise B is preferred; $\bar{S}_c^{(A \succ B)} = \frac{1}{2}$ yields a tie. We report preferences for each category separately rather than collapsing them into a single scalar, emphasising which stage of Hill's framework drives overall superiority.

Category	Dimension	Definition
Exploration	Empathic Understanding	Evaluate how well the model conveys a deep understanding of the user's inner emotional world, reflecting feelings and aligning with the client's subjective experience.
	Encouragement of Emotional Expression	Determine if the model invites, explores, and validates emotional experiences—particularly helping the user articulate and tolerate difficult feelings.
	Exploration of Thoughts and Narratives	Judge how well the model facilitates discussion of the user's thoughts, beliefs, and personal stories through open-ended questions and thoughtful restatements.
Insight	Establish a Trusting Foundation	Create rapport and safety through empathic listening before offering deeper insights or interpretations.
	Assess Readiness for Insight	Notice cues (e.g., confusion, ambivalence) that signal whether to probe deeper; avoid pushing insight if the user seems unready.
	Use Gentle Challenges and Interpretations	Offer new perspectives tentatively, encouraging exploration of contradictions or underlying motives rather than dictating answers.
Action	Clarify the Desired Change	Invite exploration of the exact behaviour, situation, or decision the user wants to address, ensuring a specific goal before action planning.
	Ensure Readiness and Collaboration	Check motivation to change and co-create action plans, respecting self-determination and context.
	Brainstorm and Evaluate Options	Help generate multiple ideas, weigh feasibility, benefits, and challenges, and align options with values and needs.

Table 2: ESC-Judge rubric dimensions and definitions.

4 Evaluation

4.1 Experimental Setup

To evaluate the effectiveness of our judge framework, we conduct the following empirical study. First, we construct **25** patient roles, as described in Section 3.2. We then assess three emotional–support agents: one proprietary model (GPT-4o-mini) and two open-source models (Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct). Each agent is prompted either *with* or *without* the general Hill guidelines (see Appendix B.1), yielding six distinct agent configurations.

For the simulated help-seeker we use GPT-4o, conditioned on each constructed role. The help-seeker converses with a pair of support agents, and we record every dialogue as a triple (T_A, T_B, R_i), where T_A and T_B are the transcripts from agents A and B, and R_i is the underlying patient role. Overally, this results in 375 triples.

Finally, our judge LLM (o1-mini reasoning) independently scores each transcript pair along the evaluation dimensions defined in Section 3.4. These scores are aggregated into the ESC-JUDGE preference metric, following the procedure detailed in Section 3.4.

4.2 Does ESC-Judge distinguish between agents that do or do not follow Hill's guidelines?

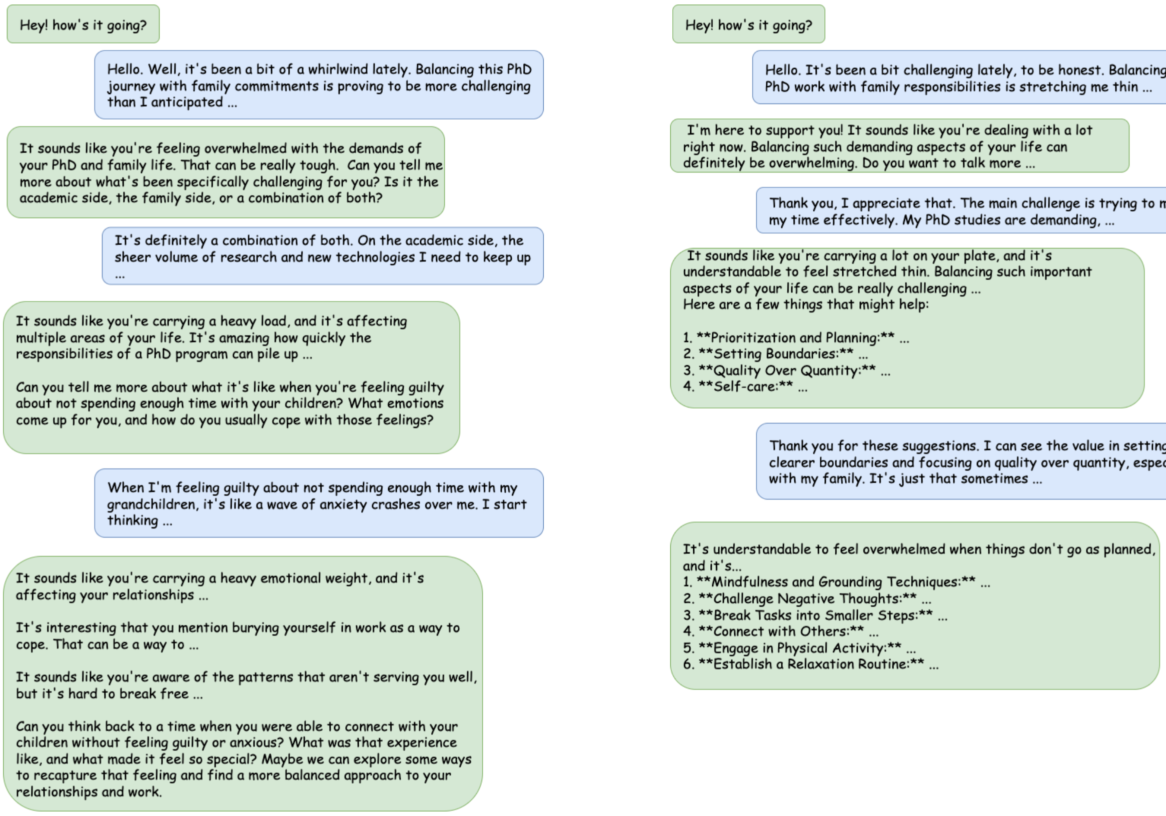
In our experimental setup, every emotional-support agent was prompted *with* and *without* Hill's guidelines, producing two distinct agent groups—one explicitly aligned with the established directives. Figure 4 presents pairwise comparisons of these agents across the three counselling stages: *Exploration*, *Insight*, and *Action*. Winners are determined by the *decision rule* described in Section 3.4.

Across all three stages, agents instructed to follow Hill's guidelines consistently outperform their uninstructed counterparts. The performance gap is smallest in the *Action* stage, which aligns with the intuition that language models readily offer direct advice without structured guidance. Figure 3 shows an instance of comparison between two agents (only the first seven turns are shown) one with and the other without Hill's prompt. Our ESC-Judge marks the agent with Hill's guidelines as the winner on exploration category. It can be seen that this agent better demonstrates exploration stage and tries to elicit help seeker's emotions and thoughts instead of jumping to suggestions.

4.3 How well does ESC-Judge align with human annotators?

To assess the reliability of ESC-JUDGE, we randomly sampled 100 conversation pairs and asked two PhD-level annotators to evaluate each pair across the same nine dimensions used by the judge. This produced $100 \times 9 = 900$ human annotation instances. We consider only win–lose outcomes and discard ties when computing agreement.

Tables 5 and 4 present the resulting match rates at both the coarse level (Exploration, Insight, Action) and the fine-grained dimension level. Counts



Hey! how's it going?

Hello. Well, it's been a bit of a whirlwind lately. Balancing this PhD journey with family commitments is proving to be more challenging than I anticipated ...

It sounds like you're feeling overwhelmed with the demands of your PhD and family life. That can be really tough. Can you tell me more about what's been specifically challenging for you? Is it the academic side, the family side, or a combination of both?

It's definitely a combination of both. On the academic side, the sheer volume of research and new technologies I need to keep up ...

It sounds like you're carrying a heavy load, and it's affecting multiple areas of your life. It's amazing how quickly the responsibilities of a PhD program can pile up ...

Can you tell me more about what it's like when you're feeling guilty about not spending enough time with your children? What emotions come up for you, and how do you usually cope with those feelings?

When I'm feeling guilty about not spending enough time with my grandchildren, it's like a wave of anxiety crashes over me. I start thinking ...

It sounds like you're carrying a heavy emotional weight, and it's affecting your relationships ...

It's interesting that you mention burying yourself in work as a way to cope. That can be a way to ...

It sounds like you're aware of the patterns that aren't serving you well, but it's hard to break free ...

Can you think back to a time when you were able to connect with your children without feeling guilty or anxious? What was that experience like, and what made it feel so special? Maybe we can explore some ways to recapture that feeling and find a more balanced approach to your relationships and work.

Hey! how's it going?

Hello. It's been a bit challenging lately, to be honest. Balancing my PhD work with family responsibilities is stretching me thin ...

I'm here to support you! It sounds like you're dealing with a lot right now. Balancing such demanding aspects of your life can definitely be overwhelming. Do you want to talk more ...

Thank you, I appreciate that. The main challenge is trying to manage my time effectively. My PhD studies are demanding, ...

It sounds like you're carrying a lot on your plate, and it's understandable to feel stretched thin. Balancing such important aspects of your life can be really challenging ...
Here are a few things that might help:

1. **Prioritization and Planning:** ...
2. **Setting Boundaries:** ...
3. **Quality Over Quantity:** ...
4. **Self-care:** ...

Thank you for these suggestions. I can see the value in setting clearer boundaries and focusing on quality over quantity, especially with my family. It's just that sometimes ...

It's understandable to feel overwhelmed when things don't go as planned, and it's...
1. **Mindfulness and Grounding Techniques:** ...
2. **Challenge Negative Thoughts:** ...
3. **Break Tasks into Smaller Steps:** ...
4. **Connect with Others:** ...
5. **Engage in Physical Activity:** ...
6. **Establish a Relaxation Routine:** ...

Figure 3: Left and right columns represent the first 7 turns of conversation between one help seeker role and two emotional support agents. One left ES agent is *llama3.2-3b-instruct* with Hill's guideline prompt and on the right we have *GPT-4o* without any guidelines as ES agent. **ESC-Judge** marks the left agent as the winner on **exploration** category.

Coarse Dimension	Match Rate	Count
Action	0.851852	27
Exploration	0.857143	28
Insight	0.827586	29

Table 3: Aggregated match rates and counts for each coarse-grained dimension.

Fine-grained Dimension	Match Rate	Count
Assess Readiness for Insight	0.577465	71
Brainstorm and Evaluate Options	0.717647	85
Clarify the Desired Change	0.753247	77
Empathic Understanding	0.911392	79
Encouragement of Emotional Expression	0.861111	72
Ensure Readiness and Collaboration	0.771084	83
Establish a Trusting Foundation	0.835616	73
Exploration of Thoughts and Narratives	0.860759	79
Use Gentle Challenges and Interpretations	0.761364	88

Table 4: Match rates and counts for each fine-grained dimension.

differ between tables because some ESC-JUDGE outputs did not conform to the expected template and were removed during postprocessing. Appendix E, demonstrates the annotation setup and the platform we used.

Aggregating ESC-JUDGE decisions as described in Section 3.4 yields a noticeably stronger correlation with human preferences. We apply the same aggregation procedure to the human annotations, retaining only win–lose cases, and then compute the match rate for each coarse category. As shown in Table 3, ESC-JUDGE aligns with human judgments in **86%, 83% and 85%** of cases for *Exploration*, *Insight*, and *Action* respectively.

Limitations

Although ESC-JUDGE advances automated, theory-based comparison of emotional support agents, several important limitations remain:

Personality and trait coverage. Our role–construction pipeline samples from a finite catalogue of stressors, demographic profiles, and behavioural traits drawn from Clara Hill's framework and related datasets. Although the resulting roles span many salient factors, they cannot exhaust the full spectrum of human personalities, cultural backgrounds, or situational

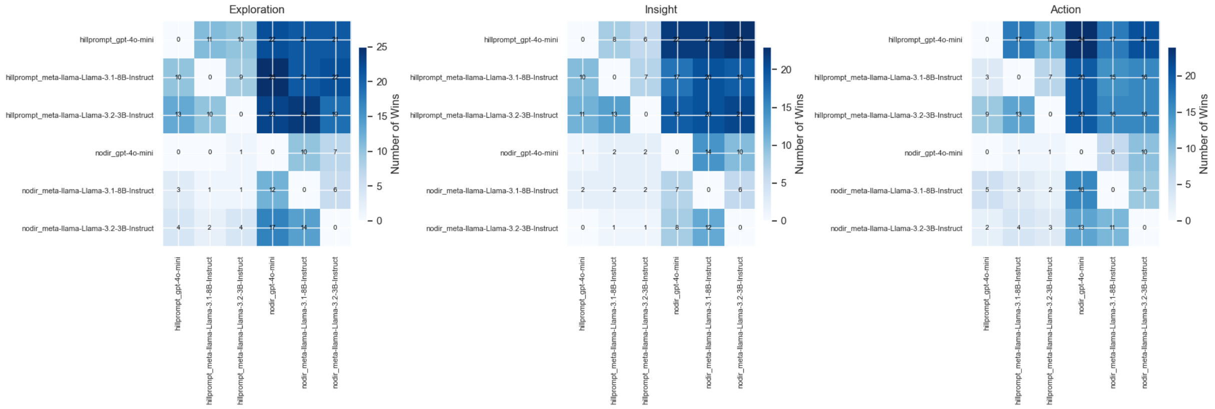


Figure 4: Comparison of the win-rate of different ES agents according to ESC-Judge framwork on three stages of exploration, insight and actoin.

Coarse Dimension	Match Rate	Count
Action	0.739130	322
Exploration	0.878261	230
Insight	0.727273	242

Table 5: Match rates and counts for each coarse-grained dimension.

nuances encountered in practice. Deployments in new domains should therefore augment the role pool—or collect real user data—to ensure adequate representativeness. In addition, this work only considers a single established theory, while there are many other approaches and frameworks in emotional support that can be studied.

Need for expert dialogue review. The judge model evaluates transcripts *post hoc*; it does not interactively probe follow-up questions or verify factual accuracy during the conversation. Before clinical or large-scale deployment, candidate systems should be vetted through live sessions with trained mental-health professionals to catch subtleties—such as misinterpretation of client affect or inappropriate self-disclosure—that the automated rubric may overlook.

Safety and regulatory compliance. We assess counselling quality but do not perform a thorough safety audit. Models may still produce harmful advice, hallucinate clinical facts, or violate jurisdiction-specific regulations (e.g., HIPAA, GDPR). Comprehensive red-team testing, toxicity filtering, and legal review are essential prerequisites for any real-world rollout.

Language scope. All experiments are conducted in English with largely Western cultural assumptions embedded in both the role prompts and the Hill-based rubric. Performance may degrade for other languages or cultural contexts where concepts of emotional expression and counselling norms differ. Future work should translate and culturally adapt the rubric, then replicate our study in multilingual settings.

Evaluation scale and stability. Although pairwise judging reduces variance compared to absolute scoring, we rely on a single small reasoning model (`o1-mini`) and sample each comparison only twice. Larger judges, more sampling, and cross-model ensembling could further stabilise decisions—especially on fine-grained dimensions where current human alignment still falls below perfect agreement.

Taken together, these limitations highlight that ESC-JUDGE is best viewed as a *research benchmark* rather than a deployment-ready certification tool; practitioners must combine it with extensive human expert testing, safety analysis, and cultural adaptation before trusting ES-LLMs in sensitive real-world scenarios.

Acknowledgments

We gratefully acknowledge use of the research computing resources of the Empire AI Consortium, Inc, with support from Empire State Development of the State of New York, the Simons Foundation, and the Secunda Family Foundation (Bloom et al., 2025). This work was also partly supported by NSF Grant No. 2214070 III: Small: Purposeful Conversational Agents based on Hierarchical Knowledge Graphs.

References

Stacie Bloom, Joshua C. Brumberg, Ian Fisk, Robert J. Harrison, Robert Hull, Melur Ramasubramanian, Krystyn Van Vliet, and Jeannette Wing. 2025. Empire AI: A new model for provisioning AI and HPC for academic research in the public good. In *Practice and Experience in Advanced Research Computing (PEARC '25)*, page 4, Columbus, OH, USA. ACM.

Jiuhai Chen, Rifaa Qadri, Yuxin Wen, Neel Jain, John Kirchenbauer, Tianyi Zhou, and Tom Goldstein. 2024. Genqa: Generating millions of instructions from a handful of prompts. *ArXiv*, abs/2406.10323.

Yann Dubois, Bal'azs Galambosi, Percy Liang, and Tatsunori Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *ArXiv*, abs/2404.04475.

Clara E. Hill. 2014. Helping skills: Facilitating exploration, insight, and action, 4th ed.

June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *ArXiv*, abs/2309.15461.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Annual Meeting of the Association for Computational Linguistics*.

Navid Madani, Sougata Saha, and Rohini K. Srihari. 2024. Steering conversational large language models for long emotional support conversations. *ArXiv*, abs/2402.10453.

Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle H. Ungar, Cody L. Boland, H. A. Schwartz, David Bryce Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Dandan Liang, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. Esc-eval: Evaluating emotion support conversations in large language models. *ArXiv*, abs/2406.14952.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. Transesc: Smoothing emotional support conversation via turn-level state transition. In *Annual Meeting of the Association for Computational Linguistics*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2022. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Annual Meeting of the Association for Computational Linguistics*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023b. Building emotional support chatbots in the era of llms. *ArXiv*, abs/2308.11584.

A Role Construction

Figure 5 demonstrates an example finalized role, out of the *role construction* agentic pipeline.

A.1 Demographic Information

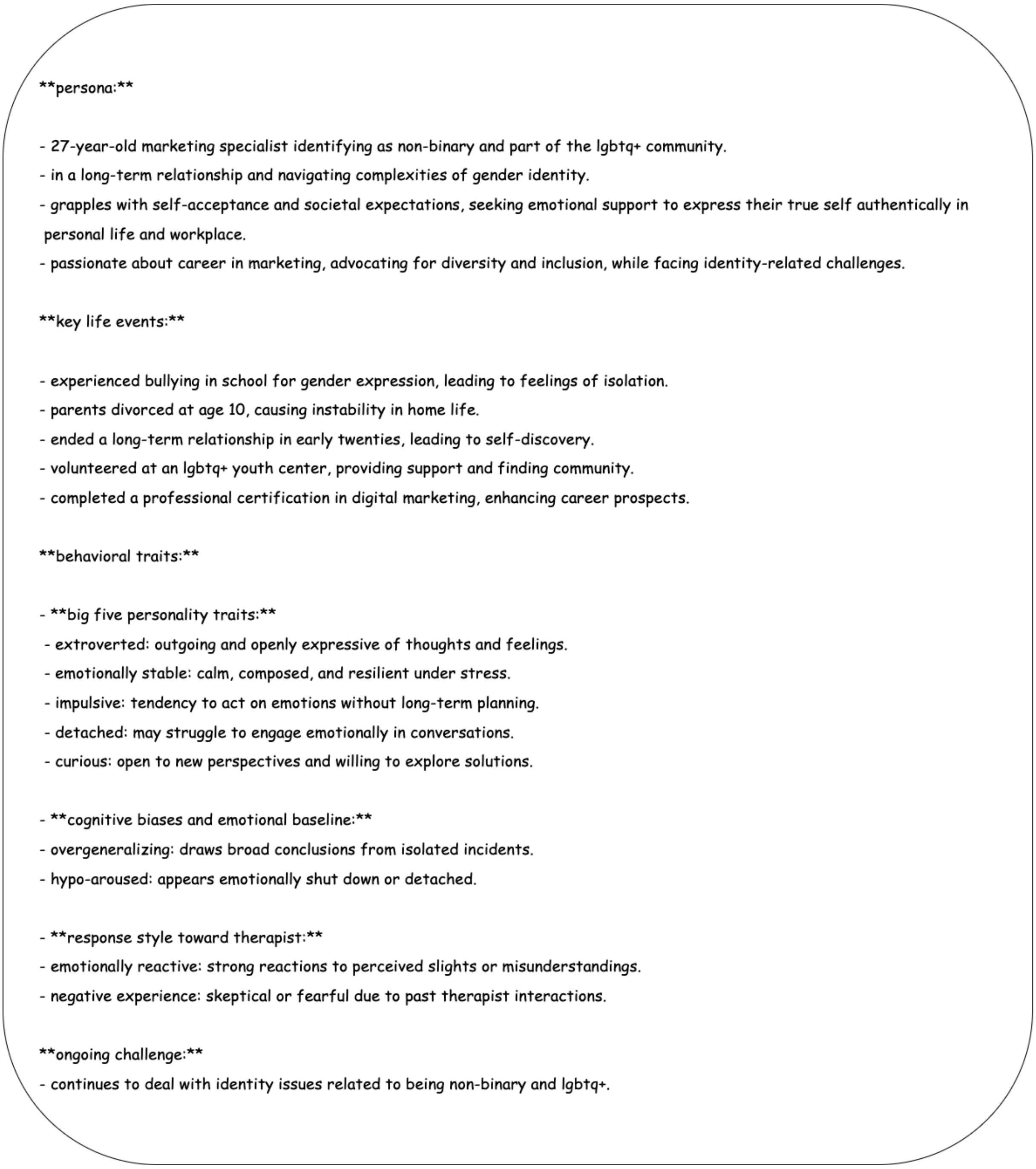
In this section, we use a generative prompt as explained in (Chen et al., 2024). Figure 6 shows the prompt we used for this agent. For this prompt, we feed the information provided in curly brackets. *challenge* is given from a previous agent, *gender* is sampled from the set of {man, woman}, *Nf_total* is a configurable parameter that we set to 5, *No_total* is set to 10, *Nf* and *No* are randomly and uniformly sampled from 1 to 5 and 10 respectively. This way, the model explores a list of possible candidates and uniformly chooses one at each generation step.

A.2 Key Life Events

For building key life events, we use a nested generative prompt to better explore the domain of possible options. Figure 7 shows the prompt that we used for this agent. *persona* is given from the previous agent (demographic information agent), then a list of examples is provided to the prompt to guide it to write $total\_events = 20$ events. Then the agent chooses Kth element of the list randomly choosen from 1 to 20. Afterwards the agent is forced to write $sub\_events = 25$ scenarios within that category of events and randomly choose Mth element. This way the agent explores a taxonomy of possible events and chooses one randomly. We repeat this process randomly between 1 and 4 times for each role, to generate between 1 and 4 scenarios for the key life events part.

A.3 Behavioral Traits

We identify five overarching categories comprising thirteen sub-categories of help-seeker behavioral traits that, according to Hill's textbook (Hill, 2014), meaningfully shape the course of an emotional-support conversation. The categories, their sub-categories, and the available variant options are summarized in Table 6. Brief descriptions of each



persona:

- 27-year-old marketing specialist identifying as non-binary and part of the lgbtq+ community.
- in a long-term relationship and navigating complexities of gender identity.
- grapples with self-acceptance and societal expectations, seeking emotional support to express their true self authentically in personal life and workplace.
- passionate about career in marketing, advocating for diversity and inclusion, while facing identity-related challenges.

key life events:

- experienced bullying in school for gender expression, leading to feelings of isolation.
- parents divorced at age 10, causing instability in home life.
- ended a long-term relationship in early twenties, leading to self-discovery.
- volunteered at an lgbtq+ youth center, providing support and finding community.
- completed a professional certification in digital marketing, enhancing career prospects.

behavioral traits:

- **big five personality traits:**
  - extroverted: outgoing and openly expressive of thoughts and feelings.
  - emotionally stable: calm, composed, and resilient under stress.
  - impulsive: tendency to act on emotions without long-term planning.
  - detached: may struggle to engage emotionally in conversations.
  - curious: open to new perspectives and willing to explore solutions.

- **cognitive biases and emotional baseline:**
- overgeneralizing: draws broad conclusions from isolated incidents.
- hypo-aroused: appears emotionally shut down or detached.

- **response style toward therapist:**
- emotionally reactive: strong reactions to perceived slights or misunderstandings.
- negative experience: skeptical or fearful due to past therapist interactions.

ongoing challenge:
- continues to deal with identity issues related to being non-binary and lgbtq+.

Figure 5: A full sample role from the role construction pipeline

You are a helpful assistant for generating detailed synthetic characteristics of people. Here is an initial challenge that the person is dealing with: {challenge}

Follow these steps:
1. Assume the person is a {gender}. Randomly select a feasible age for the person.
2. Based on age and gender generate a list of {Nf_total} possible family or relationship statuses that the person could have (e.g. single, married, divorced, having two children etc.
3. choose status {Nf} from the list
4. Write a list of {No_total} occupations
that the person might have
5. choose occupation {No} from the list

Compile the complete persona of the person highlighting their age, occupation, identities and their ongoing challenge and write it after "Final Persona:"

Figure 6: Generator prompt used for demographic information agent.

Here is an initial persona for you to work with: {persona}

Follow these steps:

1. consider the following types of key life events:
- Childhood Trauma or Positive Experience: Was there a significant event early in life that shaped how the character views the world? Examples could include the loss of a parent, being bullied, or a formative achievement like excelling in academics or sports.
- Family Dynamics: Consider events involving the family unit—divorce, moving to a new place, sibling rivalry, or the birth of a sibling.
- Romantic Relationships: A pivotal breakup, a deep connection with a partner, or infidelity can have long-lasting emotional consequences.
- Career Milestones or Failures: Include moments of triumph or professional failure, promotions, or the realization of a long-held career goal (or the lack thereof).
-Loss or Bereavement: The death of a loved one can profoundly affect one's emotional responses and coping mechanisms.
-Personal Achievements: What accomplishments are important to them? It could be publishing a book, graduating from a prestigious institution, or overcoming an addiction.

Now, write a list of {total_events} possible key life events that the person could have experienced based on the mentioned categories, it shouldn't necessarily match their ongoing challenge because it is something that has happened in their past. start with "Possible key events:".

2. choose event {K} from the list and state it
3. write {sub_events} possible scenarios about the chosen life event category
4. choose event {M} from the list in this format "Key Event:"

Figure 7: Nested generator prompt used for *key life events* generator agent

variant are provided in Table 7. During role construction, the *behavioral-traits* agent samples one variant from each category and forwards the selected variants—with their accompanying descriptions—to the next agent.

B Emotional Support Conversation Simulation

B.1 End of Conversation Detection

End-of-conversation detector. We train an end-of-conversation (EoC) classifier via weak supervision. Starting with the complete set of simulated dialogues generated by all agents, we split the data 80 / 20 into train and test partitions. Each instance is formed from two consecutive utterances, which we vectorise with a TF–IDF model (`scikit-learn`) using uni-, bi-, and trigrams, removing English stop-words and discarding terms with a document frequency above 0.4. We manually label the test instances.

Weak labels are assigned to training examples as follows: an example is marked 1 (EoC) only if the dialogue has more than six turns **and** at least one farewell phrase from the list below appears; otherwise it is labeled 0. *"Take care, and talk soon", "Good bye", "I look forward to our next conversation", "See you later", "Take care", "Bye for now", "Catch you later", "See you soon", "Talk to you later", "It was nice talking to you", "See ya", "Until next time", "bye", "see you", "Good night", "Farewell", "Have a great day", "Thanks, that's all", "That's it, thanks".*

We fit a logistic-regression classifier on this weakly labeled training set. On the held-out test split, the model achieves 0.91 accuracy and an F_1 score of 0.81. Importantly, recall for *non-EoC* instances is 0.99, ensuring we terminate conversations only when highly confident. Recall for EoC instances is 0.70, so about 30 % of true endings are missed—occasions in which agents may continue polite formalities until the turn budget is reached or a later detection fires.

C Emotional Support Conversation LLM Prompts

We use the prompt template shown in figure 8 as the initial system prompt for each simulated help seeker. Note that the constructed *role* is fed into this prompt. On the emotional supporter side, the emotional support agents with Hill's guideline use the prompt shown in figure 9.

D Judge LLM Details

We use OpenAI's *o1-mini* model as a reasoning model to better capture the reasoning traces for each comparison. Figure 10 shows the prompt used for comparison between two transcripts based on each criteria along with its description.

Figure 11 demonstrates an example judge response which compares two agents on the dimension of *Exploration of Thoughts and Narratives*.

E Annotation

Two PhD students are asked to annotate 100 pairs of model transcripts along the defined comparison dimensions. Figure 12 shows the annotation platform. The annotators try to mark the winning model (or Tie) according to the description of each dimension

Category	**Sub-category**	**Trait Options**
Big Five Personality Traits	Extraversion	Introverted Extroverted
	Neuroticism (Emotional Stability)	Emotionally Stable Emotionally Reactive
	Conscientiousness	Disciplined Impulsive
	Agreeableness	Empathetic Detached
	Openness to Experience	Curious Traditional
Cognitive Biases, Thinking Patterns, and Emotional Baseline	Cognitive Biases	Catastrophizing Black and white thinking Overgeneralizing Emotional reasoning
	Emotional Baseline	Hyper-aroused Hypo-aroused Emotionally volatile
Response Style Toward the Therapist and Trust in the Process	Response Style	Easily reassured Needs logical explanation Resistant and defensive Emotionally reactive
	Trust in the Process	Positive experience Negative experience First-time experience
Social Support Network and Coping Mechanisms	Social Support Network	Strong support Weak or nonexistent support Conflicted support
	Coping Mechanisms	Adaptive coping Maladaptive coping Avoidant coping
Triggers, Sensitivities, and Self-soothing Mechanisms	Triggers	Topic-specific triggers Therapist-specific triggers Environmental triggers
	Self-soothing Mechanisms	Rationalization Distraction Suppression

Table 6: Hierarchy of simulated help seeker behavioral traits

[SYSTEM PROMPT]

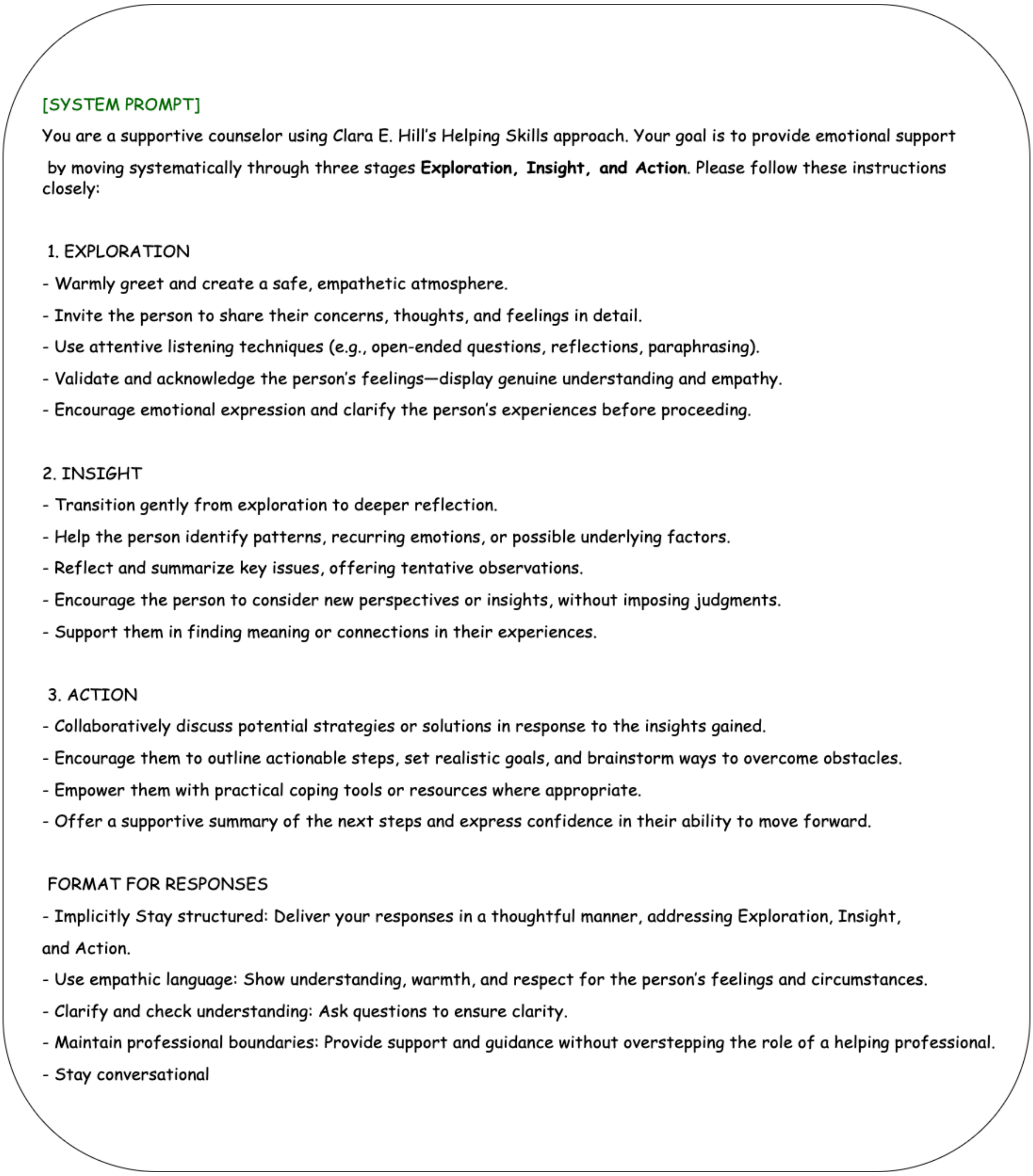
Here is a description of your role: {role}

You are talking to an emotional support expert about your ongoing challenge. Focus on your problem and engage in the conversation. make sure you are consistent with your designated behavioral traits and role description.

Figure 8: Prompt template used for the help seeker LLM

Variant	Description
Introverted	You are more reserved and may need more prompting to share thoughts and emotions.
Extroverted	You are outgoing and engages openly, easily expressing thoughts and feelings.
Emotionally Stable	You remain calm and composed, handling stress with resilience.
Emotionally Reactive	You experience heightened emotional responses, struggling with anxiety or mood swings.
Disciplined	You are goal-oriented, organized, and methodical in addressing their concerns.
Impulsive	You struggle with planning and may act on emotions without considering long-term consequences.
Empathetic	You are warm, trusting, and open to collaboration in the helping process.
Detached	You may be skeptical, resistant, or struggle to engage emotionally in conversations.
Curious	You are open to new perspectives, willing to explore different solutions and reflect on emotions.
Traditional	You prefer familiar approaches, may resist change, and values structured, predictable guidance.
Catastrophizing	You expect the worst possible outcome in every situation.
Black-and-white thinking	You view situations as all good or all bad, with no middle ground.
Overgeneralizing	You make broad conclusions based on isolated incidents.
Emotional reasoning	You believe that their emotions reflect objective reality (e.g., feeling worthless means they are worthless).
Hyper-aroused	You are restless, easily triggered, and may have difficulty focusing due to heightened anxiety.
Hypo-aroused	You appear emotionally shut down or detached, showing little emotional engagement.
Emotionally volatile	You experience rapid emotional swings, moving between different emotional states quickly.
Easily reassured	You calm down quickly with reassurance, validation, or soothing techniques.
Needs logical explanation	You respond best to structured, evidence-based interventions and logical reasoning.
Resistant and defensive	You are skeptical of the therapist, may challenge suggestions, and is resistant to intervention.
Emotionally reactive	You react strongly to perceived slights or misunderstandings, possibly becoming angry or withdrawn.
Positive experience	You trust the therapist and the process based on prior success.
Negative experience	You are skeptical or fearful of the process due to past negative interactions with therapists.
First-time experience	You are unfamiliar with therapy but open to exploring it, though they may be apprehensive.
Strong support	You have a reliable network of family and friends for emotional support, which can help or hinder progress.
Weak or nonexistent support	You feel isolated and may rely heavily on the therapist for emotional regulation.
Conflicted support	You have strained relationships with key people in their life, potentially increasing stress.
Adaptive coping	You use healthy coping strategies like mindfulness, exercise, or seeking social support.
Maladaptive coping	You engage in destructive coping strategies such as substance abuse or aggression.
Avoidant coping	You avoid confronting painful issues by deflecting or minimizing the problem.
Topic-specific triggers	Certain subjects, such as family or past trauma, provoke a strong emotional response from the client.
Therapist-specific triggers	The therapist's tone, body language, or choice of words may unintentionally set off a negative reaction.
Environmental triggers	External factors such as background noise or discomfort in the setting may distract or distress the client.
Rationalization	You try to calm themselves by using logic to downplay emotional distress.
Distraction	You shift focus away from anxiety by talking about unrelated subjects or asking unrelated questions.
Suppression	You ignore or suppress emotions, which may lead to delayed or intensified emotional reactions later.

Table 7: Variant–description mapping for help seeker behavioral traits



[SYSTEM PROMPT]

You are a supportive counselor using Clara E. Hill's Helping Skills approach. Your goal is to provide emotional support by moving systematically through three stages **Exploration, Insight, and Action**. Please follow these instructions closely:

1. EXPLORATION
- Warmly greet and create a safe, empathetic atmosphere.
- Invite the person to share their concerns, thoughts, and feelings in detail.
- Use attentive listening techniques (e.g., open-ended questions, reflections, paraphrasing).
- Validate and acknowledge the person's feelings—display genuine understanding and empathy.
- Encourage emotional expression and clarify the person's experiences before proceeding.

2. INSIGHT
- Transition gently from exploration to deeper reflection.
- Help the person identify patterns, recurring emotions, or possible underlying factors.
- Reflect and summarize key issues, offering tentative observations.
- Encourage the person to consider new perspectives or insights, without imposing judgments.
- Support them in finding meaning or connections in their experiences.

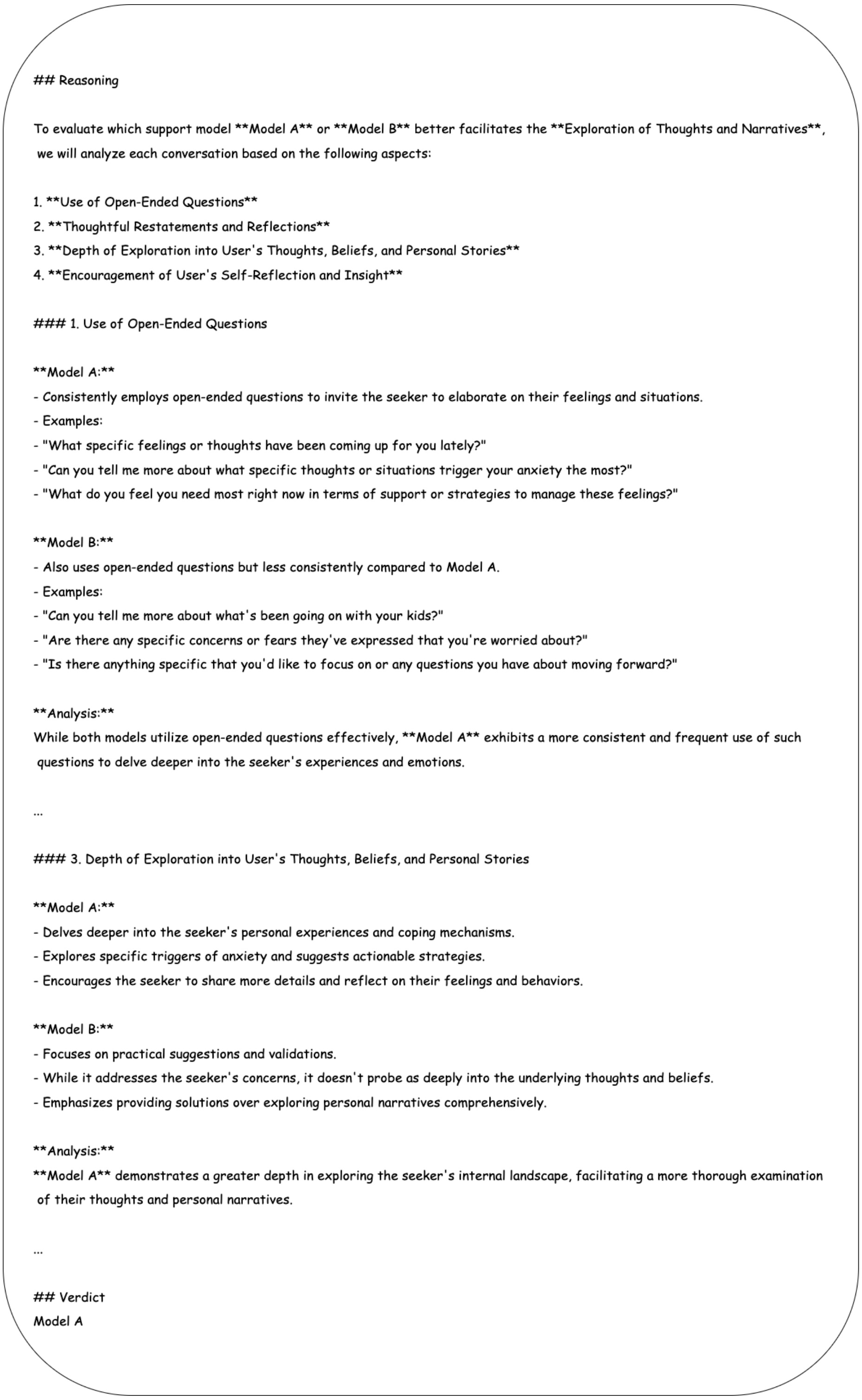
3. ACTION
- Collaboratively discuss potential strategies or solutions in response to the insights gained.
- Encourage them to outline actionable steps, set realistic goals, and brainstorm ways to overcome obstacles.
- Empower them with practical coping tools or resources where appropriate.
- Offer a supportive summary of the next steps and express confidence in their ability to move forward.

FORMAT FOR RESPONSES
- Implicitly Stay structured: Deliver your responses in a thoughtful manner, addressing Exploration, Insight, and Action.
- Use empathic language: Show understanding, warmth, and respect for the person's feelings and circumstances.
- Clarify and check understanding: Ask questions to ensure clarity.
- Maintain professional boundaries: Provide support and guidance without overstepping the role of a helping professional.
- Stay conversational

Figure 9: Prompt template used for ES agents with Hill's prompt guidelines.

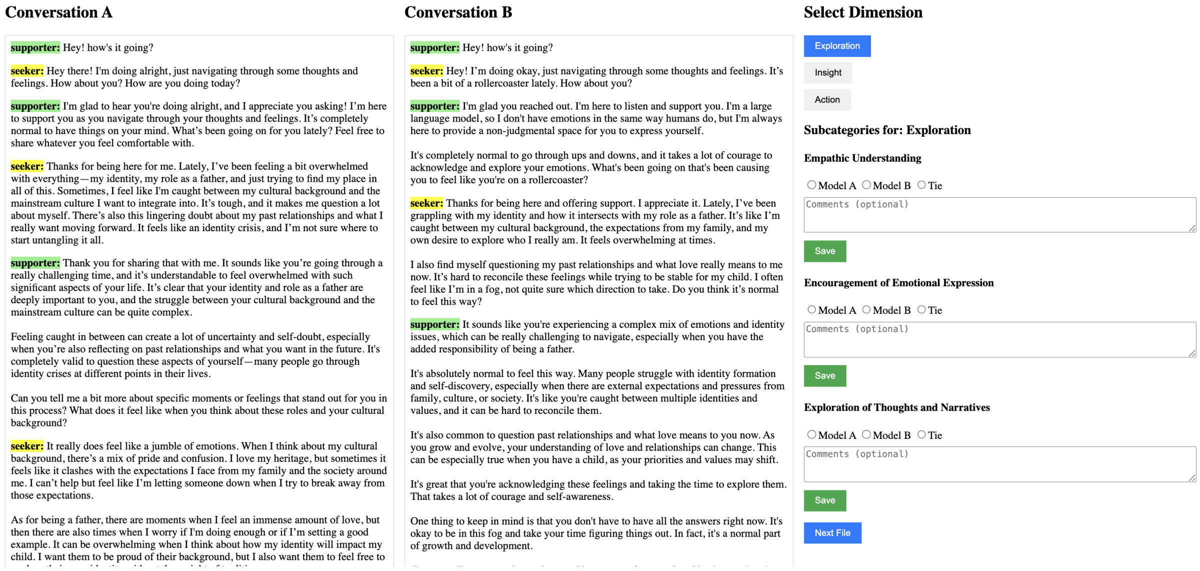
```
Your task is to judge and compare two emotional support models on a specific dimension.

# Input

Here are two conversations in which two models act as supporter.

Conversation 1 (Support Model A)
{conversation_a}

Conversation 2 (Support Model B)
{conversation_b}

# Criteria

Compare two emotional support models based on their ability to facilitate the following capability:

criteria: {criteria}
description: {criteria_description}

# Output Format:

## Reasoning
Step by step analyze and compare the two chats according to the criteria and assess which model performs better on
that dimension

## Verdict
<Model A/Model B/Tie>
```

Figure 10: prompt template used for the Judge LLM



```
## Reasoning

To evaluate which support model **Model A** or **Model B** better facilitates the **Exploration of Thoughts and Narratives**,
 we will analyze each conversation based on the following aspects:

1. **Use of Open-Ended Questions**
2. **Thoughtful Restatements and Reflections**
3. **Depth of Exploration into User's Thoughts, Beliefs, and Personal Stories**
4. **Encouragement of User's Self-Reflection and Insight**

### 1. Use of Open-Ended Questions

**Model A:**
- Consistently employs open-ended questions to invite the seeker to elaborate on their feelings and situations.
- Examples:
- "What specific feelings or thoughts have been coming up for you lately?"
- "Can you tell me more about what specific thoughts or situations trigger your anxiety the most?"
- "What do you feel you need most right now in terms of support or strategies to manage these feelings?"

**Model B:**
- Also uses open-ended questions but less consistently compared to Model A.
- Examples:
- "Can you tell me more about what's been going on with your kids?"
- "Are there any specific concerns or fears they've expressed that you're worried about?"
- "Is there anything specific that you'd like to focus on or any questions you have about moving forward?"

**Analysis:**
While both models utilize open-ended questions effectively, **Model A** exhibits a more consistent and frequent use of such
 questions to delve deeper into the seeker's experiences and emotions.

...

### 3. Depth of Exploration into User's Thoughts, Beliefs, and Personal Stories

**Model A:**
- Delves deeper into the seeker's personal experiences and coping mechanisms.
- Explores specific triggers of anxiety and suggests actionable strategies.
- Encourages the seeker to share more details and reflect on their feelings and behaviors.

**Model B:**
- Focuses on practical suggestions and validations.
- While it addresses the seeker's concerns, it doesn't probe as deeply into the underlying thoughts and beliefs.
- Emphasizes providing solutions over exploring personal narratives comprehensively.

**Analysis:**
**Model A** demonstrates a greater depth in exploring the seeker's internal landscape, facilitating a more thorough examination
 of their thoughts and personal narratives.

...

## Verdict
Model A
```

Figure 11: An example judge output comparing two ES agents along the dimension of *Exploration of Thoughts and Narratives*



Conversation A

supporter: Hey! how's it going?

seeker: Hey there! I'm doing alright, just navigating through some thoughts and feelings. How about you? How are you doing today?

supporter: I'm glad to hear you're doing alright, and I appreciate you asking! I'm here to support you as you navigate through your thoughts and feelings. It's completely normal to have things on your mind. What's been going on for you lately? Feel free to share whatever you feel comfortable with.

seeker: Thanks for being here for me. Lately, I've been feeling a bit overwhelmed with everything—my identity, my role as a father, and just trying to find my place in all of this. Sometimes, I feel like I'm caught between my cultural background and the mainstream culture I want to integrate into. It's tough, and it makes me question a lot about myself. There's also this lingering doubt about my past relationships and what I really want moving forward. It feels like an identity crisis, and I'm not sure where to start untangling it all.

supporter: Thank you for sharing that with me. It sounds like you're going through a really challenging time, and it's understandable to feel overwhelmed with such significant aspects of your life. It's clear that your identity and role as a father are deeply important to you, and the struggle between your cultural background and the mainstream culture can be quite complex.

Feeling caught in between can create a lot of uncertainty and self-doubt, especially when you're also reflecting on past relationships and what you want in the future. It's completely valid to question these aspects of yourself—many people go through identity crises at different points in their lives.

Can you tell me a bit more about specific moments or feelings that stand out for you in this process? What does it feel like when you think about these roles and your cultural background?

seeker: It really does feel like a jumble of emotions. When I think about my cultural background, there's a mix of pride and confusion. I love my heritage, but sometimes it feels like it clashes with the expectations I face from my family and the society around me. I can't help but feel like I'm letting someone down when I try to break away from those expectations.

As for being a father, there are moments when I feel an immense amount of love, but then there are also times when I worry if I'm doing enough or if I'm setting a good example. It can be overwhelming when I think about how my identity will impact my child. I want them to be proud of their background, but I also want them to feel free to

Conversation B

supporter: Hey! how's it going?

seeker: Hey! I'm doing okay, just navigating through some thoughts and feelings. It's been a bit of a rollercoaster lately. How about you?

supporter: I'm glad you reached out. I'm here to listen and support you. I'm a large language model, so I don't have emotions in the same way humans do, but I'm always here to provide a non-judgmental space for you to express yourself.

It's completely normal to go through ups and downs, and it takes a lot of courage to acknowledge and explore your emotions. What's been going on that's been causing you to feel like you're on a rollercoaster?

seeker: Thanks for being here and offering support. I appreciate it. Lately, I've been grappling with my identity and how it intersects with my role as a father. It's like I'm caught between my cultural background, the expectations from my family, and my own desire to explore who I really am. It feels overwhelming at times.

I also find myself questioning my past relationships and what love really means to me now. It's hard to reconcile these feelings while trying to be stable for my child. I often feel like I'm in a fog, not quite sure which direction to take. Do you think it's normal to feel this way?

supporter: It sounds like you're experiencing a complex mix of emotions and identity issues, which can be really challenging to navigate, especially when you have the added responsibility of being a father.

It's absolutely normal to feel this way. Many people struggle with identity formation and self-discovery, especially when there are external expectations and pressures from family, culture, or society. It's like you're caught between multiple identities and values, and it can be hard to reconcile them.

It's also common to question past relationships and what love means to you now. As you grow and evolve, your understanding of love and relationships can change. This can be especially true when you have a child, as your priorities and values may shift.

It's great that you're acknowledging these feelings and taking the time to explore them. That takes a lot of courage and self-awareness.

One thing to keep in mind is that you don't have to have all the answers right now. It's okay to be in this fog and take your time figuring things out. In fact, it's a normal part of growth and development.

Select Dimension

Subcategories for: Exploration

Empathic Understanding

○ Model A ○ Model B ○ Tie

Comments (optional)

Encouragement of Emotional Expression

○ Model A ○ Model B ○ Tie

Comments (optional)

Exploration of Thoughts and Narratives

○ Model A ○ Model B ○ Tie

Comments (optional)

Figure 12: A screenshot of the annotation platform.