Stress-Testing Emotional Support Models: Moving from Homogeneous to Diverse Help Seekers

Chaewon Heo Cheyon Jin Yohan Jo*
Graduate School of Data Science, Seoul National University
{heorshey99, cheyonjin, yohan.jo}@snu.ac.kr

Abstract

As emotional support chatbots have recently gained significant traction across both research and industry, a common evaluation strategy has emerged: use help-seeker simulators to interact with supporter chatbots. However, current simulators suffer from two critical limitations: (1) they fail to capture the behavioral diversity of real-world seekers, often portraying them as overly cooperative, and (2) they lack the controllability required to simulate specific seeker profiles. To address these challenges, we present a controllable seeker simulator driven by nine psychological and linguistic features that underpin seeker behavior. Using authentic Reddit conversations, we train our model via a **Mixture-of-Experts (MoE)** architecture, which effectively differentiates diverse seeker behaviors into specialized parameter subspaces, thereby enhancing fine-grained controllability. Our simulator achieves superior profile adherence and behavioral diversity compared to existing approaches. Furthermore, evaluating 7 prominent supporter models with our system uncovers previously obscured performance degradations. These findings underscore the utility of our framework in providing a more faithful and stress-tested evaluation for emotional support chatbots.[1]

1 Introduction

Emotional support is essential for alleviating psychological distress and maintaining mental well-being. While LLM-based emotional support systems have advanced rapidly (Zheng et al., 2025), the lack of reliable automated evaluation frameworks remains a critical bottleneck, hindering the field's establishment as an objective and measurable discipline.

Currently, the most prevalent evaluation strategy uses **help-seeker simulators** to interact with supporter models, and the resulting dialogues are assessed based on metrics such as empathy and fluency (Zhao et al., 2024). However, existing seeker simulators compromise evaluation validity due to two critical limitations. First, they fail to reflect the diversity of real-world seekers, representing only a narrow subset of seeker behaviors. Existing simulators primarily portray overly cooperative, "easy" seekers, failing to capture the vast spectrum of real-world behaviors such as advice resistance (Yaman, 2021) or limited disclosure. Second, existing simulators lack fine-grained controllability. A robust evaluation framework must allow researchers to target specific seeker populations, as this enables the identification of which aspects of a supporter model fail for particular seeker groups, providing actionable signals for model improvement.

To address these shortcomings, we present a highly controllable seeker simulator designed to take specific seeker profiles and simulate their corresponding behaviors. As a foundational component, we define the **seeker profile** as a combination of nine psychological and linguistic features such as resistance level and verbosity level (Step 1 in Figure 1). To faithfully reflect a broad array of real-world behaviors, we leverage data from Reddit online support groups as the primary training source for our simulator.

To effectively control a seeker profile defined by nine features, we employ a Mixture-of-Experts (MoE) architecture. This architecture dynamically assigns weights to specialized adapters based on the input seeker profile, accurately reflecting distinct behavioral patterns across seeker types.

Validation results highlight the effectiveness of our simulator across three key dimensions: (1) **superior profile adherence**, successfully maintaining consistent seeker traits throughout the interaction; (2) **high behavioral diversity**, capturing a wider array of help-seeking patterns compared to existing baselines; and (3) **higher fidelity**, as

*Corresponding author

[1]Our code and data are available at https://github.com/holi-lab/ES-Evaluation-Framework.

Findings of the Association for Computational Linguistics: ACL 2026, pages 22842–22869

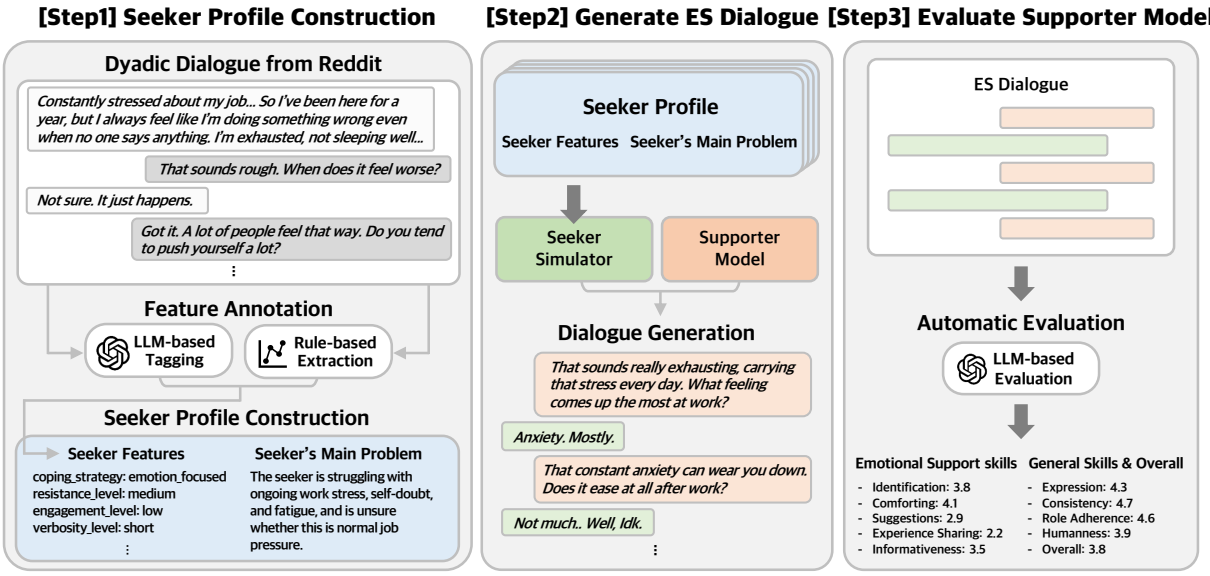


Figure 1: Overview of our evaluation framework for emotional support models.

confirmed by expert evaluations indicating that the generated responses more closely resemble real-world emotional support interactions.

Using our seeker simulator, we evaluate supporter models through interactive multi-turn dialogues, followed by automatic assessment using emotional support and general conversational skill metrics (Steps 2, 3 in Figure 1). Results indicate that supporter models exhibit varying levels of performance degradation across various seeker populations. Our framework allows researchers to flexibly define seeker populations by selecting appropriate combinations of profile features, enabling targeted and diagnostic evaluation under diverse scenarios.

In summary, we introduce a controllable seeker simulator driven by multi-dimensional seeker profiles, providing a population-diverse evaluation framework for assessing supporter models.

2 Related Works

2.1 Evaluation of Emotional Support Models

The evaluation of emotional support models has evolved from surface-level fluency to interactional supportive efficacy. Early research evaluated whether generated empathetic responses exhibited emotionally appropriate expressions (Santhanam and Shaikh, 2019), using lexical metrics (e.g., BLEU, Perplexity) against reference dialogues such as EmpatheticDialogues (Rashkin et al., 2019). Since these metrics often fail to capture supportive effectiveness, later work adopted ESConv (Liu et al., 2021) as evaluation scenarios and performed human evaluation, where evaluators role-play seekers and rate support quality using predefined criteria (e.g., Identification, Empathy) (Zhao et al., 2023; Cheng et al., 2022). However, these human role-play evaluations are constrained by predefined scenarios, fail to reflect real-world interaction dynamics, and are costly to scale. These limitations have driven recent evaluation toward simulator-based evaluation frameworks, where seeker simulators interact with supporter models and automated evaluation is performed on the generated dialogues.

2.2 Seeker Simulations in Emotional Support

Prior studies have adopted diverse approaches to implement seeker simulators. Prompt-based simulators are given input including specific characteristics, such as Big Five personality traits, or situational mental health contexts (Madani and Srihari, 2025; Qiu and Lan, 2024). Beyond simple prompting, some work employs explicit state tracking or memory modules to regulate the seeker's internal states (Yang et al., 2025; Wang et al., 2025). Another work trains simulators on emotional support datasets using profile cards with demographic attributes, enabling it to learn seeker behaviors associated with each profile (Zhao et al., 2024).

Despite conditioning on different seeker prompts, existing simulators inherit the base language model's verbose and cooperative interaction bias. Moreover, even learned simulators tend to

reflect only surface-level profile attributes, failing to induce deeper psychological traits across seeker types. Our simulator addresses these limitations.

3 Data Construction and Feature Formulation

To build a population-diverse and controllable seeker simulator, we construct a large-scale dialogue dataset grounded in authentic online interactions as the training foundation.

3.1 Data Collection and Preprocessing

We curated a dataset of real-world emotional support interactions from Reddit, specifically targeting subreddits with over 500,000 members, such as `r/offmychest` and `r/mentalhealth`. Reddit serves as an ideal source for training a feature-controllable simulator, as its in-the-wild nature encompasses a broad spectrum of informal and raw human expressions often absent in curated datasets.

Unlike traditional approaches that rely on heavily filtered datasets, we intentionally retained informal and raw elements. We included a wide variance in utterance lengths and aggressiveness, addressing the limitations of prior simulators that portray only articulate and cooperative seekers.

While prioritizing realism, we still implemented rigorous steps to ensure ethical integrity and basic data quality. Most importantly, all personally identifiable information (PII) was systematically masked to protect user anonymity. We also discarded low-quality data based on conversation length, topics, and upvote counts. The final dataset includes 11,066 dialogues. Detailed preprocessing procedures are provided in Appendix A.

3.2 Feature Taxonomy for Seeker Profiles

To represent the multifaceted characteristics of seekers' conversational behavior, we define a comprehensive taxonomy consisting of nine distinct features. These are categorized into psychological and linguistic attributes. Detailed category definitions for each feature are provided in Appendix B.

3.2.1 Psychological Features

We incorporate six psychological features to model seeker behavior. The **main coping strategy** categorizes seekers into four types, grounded in the multidimensional coping framework (Endler and Parker, 1994): *problem_focused*, *emotion_focused*, *avoidant*, and *maladaptive_behavior*. The **utterance style** captures stylistic variation in seeker expression using three categories (*plain*, *upset*, *verbose*), adapted from the PATIENT-Ψ framework (Wang et al., 2024b). To model interactional dynamics, we define a three-level ordinal **resistance level** based on psychological reactance theory (Brehm and Brehm, 1981), and a three-level **engagement level** grounded in prior work on client participation in therapeutic settings (Holdsworth et al., 2014). Finally, we include a four-level **self-disclosure level** derived from Social Penetration Theory (Altman and Taylor, 1973), as well as **seeker reaction proportions** following prior work on modeling client reactions in emotional support dialogues (Li et al., 2023).

3.2.2 Linguistic Features

We further define three linguistic attributes that capture diverse structural patterns. The **verbosity level** is defined as a five-level feature based on the distribution of seeker utterance lengths. We additionally include a binary **user profanity flag**, detected using the `profanity_check` library. Finally, the **total dialogue turns level** captures the overall length of the interaction.

3.3 Feature Annotation and Validation

To apply the feature taxonomy to our corpus, we employed a hybrid annotation framework combining LLM-based tagging and rule-based extraction.

3.3.1 Psychological Feature Annotation

Psychological features are annotated using an LLM-based tagging process, followed by targeted human validation on a subset of samples.

LLM-based Tagging Process We performed tagging at two different levels (dialogue and utterance level) and subsequently aggregated the annotations into dialogue-level representations. Detailed assignments are provided in Appendix C.

Human Validation To verify the reliability of the LLM-based feature tagging, we conducted a validation study on a sample of 60 dialogues. Inter-annotator agreement (percent agreement) averages 0.57 across psychological features, while human–LLM alignment achieves 0.84 accuracy, indicating reliable alignment. Detailed validation procedures and feature-level validation results are provided in Appendix D.

3.3.2 Linguistic Feature Extraction

Linguistic features describe the measurable characteristics of the dialogue and are extracted using rule-

based methods: *verbosity level* from seeker utterance length, *profanity flag* via `profanity-check`, and *total dialogue turns level* from turn count. Detailed extraction rules are in Appendix E.

3.4 Seeker Profile Construction

We construct a final seeker profile by combining all annotated features with the seeker's main problem. Formally, a seeker profile is represented as:

$$P = [f_1, f_2, \ldots, f_9, m],$$

where each f_i denotes an annotated **seeker feature** and m represents the **seeker's main problem**, both expressed in natural language. An example profile is in Figure 8. The resulting dataset consists of 11,066 unique seeker profiles, split into 8,868 training, 1,094 validation, and 1,104 test profiles. The seeker's main problem is a summary of the first seeker utterance (i.e., original post) that describes the seeker's situation. The summaries are generated by GPT-4o-mini (OpenAI, 2024), with detailed prompts provided in Table 12.

4 Seeker Simulator Training

To simulate diverse seeker behaviors in emotional support conversations, we employ a two-stage training framework: (1) **Supervised Fine-Tuning (SFT)** to establish a general conversational backbone, and (2) **Mixture-of-Experts (MoE) training** to enable controllable, attribute-specific generation.

4.1 Supervised Fine-Tuning

The first stage establishes a linguistic foundation by capturing general conversational patterns of seekers. We use Llama-3-8B-Instruct as the base model and apply LoRA ($r = 16$) to all linear layers. The model is trained to predict the next seeker utterance, given a seeker profile and dialogue history as a next-token prediction task. Upon convergence, the LoRA adapters are merged into the base model weights. This merged model is then entirely frozen, serving as a stable backbone for the subsequent expert-based training stage.

4.2 MoE Training with Behavioral Routing

SFT backbone often fails to manifest seeker behaviors faithful to the input seeker features since the feature signals are frequently overshadowed by long dialogue histories or system prompts.

To address this challenge, we introduce an explicit routing mechanism that decouples feature control from language modeling. Rather than expecting the model to infer seeker features from text, we provide a structured feature vector to control expert selection through a soft MoE formulation. An overview of the proposed framework is in Figure 2.

4.2.1 Seeker Feature Vector Construction

The routing network operates on a rule-based seeker feature vector $\mathbf{f} \in \mathbb{R}^{14}$ that encodes features associated with each seeker profile. This feature vector is constructed independently of the language modeling input and is used exclusively for routing, ensuring that expert selection is guided by structured behavioral signals.

Categorical attributes (e.g., *main coping strategy*) are mapped to one-hot representations for independent representation. Level-based attributes (e.g., *resistance level*, *engagement level*) are encoded as zero-centered normalized scalars to preserve their ordinal structure while maintaining comparable magnitudes across features.

4.2.2 Shared Dialogue-Level Routing Network

Given the seeker feature vector $\mathbf{f}$, we compute a routing distribution that determines how strongly each expert is activated.

The routing network consists of L residual MLP blocks followed by a gating layer. Each block applies a standard residual update, $\mathbf{h}_{\ell+1} = \mathrm{LayerNorm}(\mathbf{h}_\ell + \mathrm{MLP}(\mathbf{h}_\ell))$, and the final hidden vector $\mathbf{h}_L$ is mapped to

$$\boldsymbol{\alpha} = \mathrm{softmax}(W_g \mathbf{h}_L)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{N_E}$ and N_E is the number of experts (we set $N_E = 4$ and $L = 3$). As shown in Figure 2, $\boldsymbol{\alpha}$ is computed once per dialogue session and shared across all layers. This global routing strategy ensures that behavioral differences between seekers are reflected consistently in expert activation throughout the entire generation process.

4.2.3 LoRA-Based Experts

Using the frozen SFT backbone from Section 4.1, we attach N_E low-rank expert adapters to each linear layer in both the attention and feed-forward (FFN) blocks as illustrated in Figure 2. For an expert-augmented linear layer ℓ, the output is computed as:

$$\mathbf{y}^{(\ell)} = W^{(\ell)}\mathbf{x}^{(\ell)} + \sum_{i=1}^{N_E} \alpha_i \cdot \Delta_i^{(\ell)}(\mathbf{x}^{(\ell)}),$$

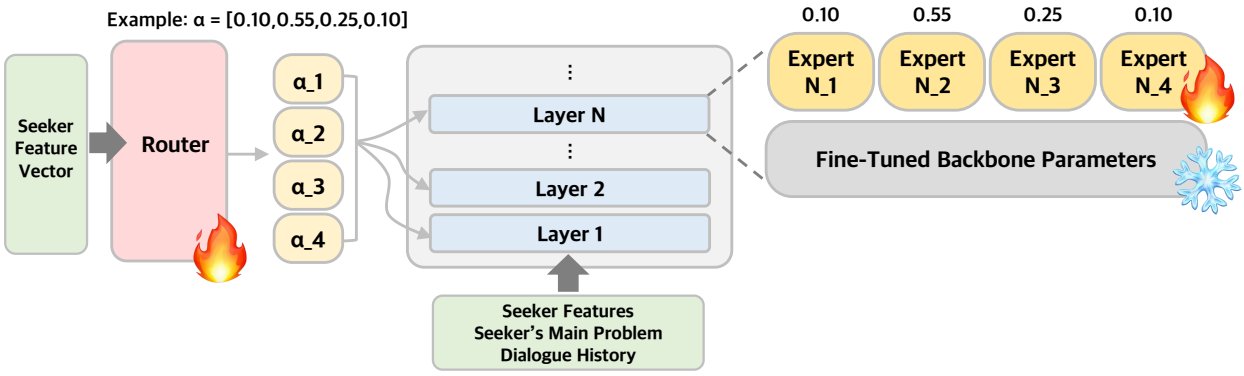


Figure 2: Overview of the MoE architecture. A frozen SFT backbone is augmented with multiple low-rank expert adapters at each linear layer. A shared routing network produces a dialogue-level routing vector $\boldsymbol{\alpha}$, which controls expert activation consistently across all layers.

where $W^{(\ell)} \in \mathbb{R}^{d_{\text{out}}^{(\ell)} \times d_{\text{in}}^{(\ell)}}$ denotes the frozen backbone weight of layer ℓ. $\Delta_i^{(\ell)}(\mathbf{x}^{(\ell)})$ is the ith expert in layer l, parameterized as a LoRA adapter, i.e., $\mathbf{x}^{(\ell)} A_i^{(\ell)} B_i^{(\ell)}$, where $A_i^{(\ell)} \in \mathbb{R}^{d_{\text{in}}^{(\ell)} \times r}$ and $B_i^{(\ell)} \in \mathbb{R}^{r \times d_{\text{out}}^{(\ell)}}$, with $r \ll d_{\text{in}}^{(\ell)}, d_{\text{out}}^{(\ell)}$ (we set $r = 4$).

4.2.4 Training Objectives

The MoE model is trained using a combination of the standard language modeling loss and a Task-wise Decorrelation (TwD) loss, following the prior work of Zhou et al. (2024).

Task-wise Decorrelation Loss. To ensure that routing distributions reflect meaningful behavioral distinctions, we encourage samples with different feature labels to produce distinct routing vectors. For each sample i, the routing distribution $\boldsymbol{\alpha}^{(i)} \in \mathbb{R}^E$ is projected into a latent space via a learnable projection function:

$$\boldsymbol{\omega}^{(i)} = \phi(\boldsymbol{\alpha}^{(i)}),$$

where $\phi(\cdot)$ denotes a learnable shared two-layer MLP that maps $\boldsymbol{\alpha} \in \mathbb{R}^{N_E}$ to a latent embedding $\boldsymbol{\omega} \in \mathbb{R}^D$, with $D = 64$ and a GELU nonlinearity.

We then optimize the routing representations using a contrastive objective defined for each feature f:

$$\mathcal{L}_{\text{TwD}}^{(f)} = \mathbb{E}_i \left[-\log \frac{\sum_{j \in \mathcal{P}_f(i)} \exp(s_{ij})}{\sum_{k \neq i} \exp(s_{ik})} \right],$$

where $\mathcal{P}_f(i) = \{ j \neq i \mid y_f^{(j)} = y_f^{(i)} \}$ denotes the set of positive samples sharing the same label for feature f as sample i, $y_f^{(i)}$ is the discrete label of sample i for feature f, and $s_{ij} = \boldsymbol{\omega}^{(i)} \cdot \boldsymbol{\omega}^{(j)} / \tau$ denotes cosine similarity with temperature τ. The final TwD loss is obtained by averaging $\mathcal{L}_{\text{TwD}}^{(f)}$ across all considered features.

Language Modeling Loss. We use a standard next-token prediction loss $\mathcal{L}_{\text{LM}}$, which is computed exclusively on the next seeker utterance from seeker profile and dialogue history, following the same setup as in Section 4.1.

Overall Objective. The final training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{TwD}}.$$

Notably, we do not pre-assign semantic roles to individual experts. Instead, expert specialization emerges implicitly through joint optimization of the routing network and expert parameters. We analyze the resulting expert behaviors and routing patterns in Section 5.1.2.

5 Seeker Simulator Validation

We evaluate our seeker simulator across three dimensions for diverse and controllable simulation. Specifically, we evaluate **(1) profile adherence**: whether the generated utterances faithfully follow the intended seeker profiles; **(2) fidelity**: how well the simulator produces human-like and psychologically plausible utterances; and **(3) diversity**: whether the simulator can cover a broad, well-separated range of seeker populations. To isolate the performance of the seeker simulator, we fix the supporter model as GPT-5-mini, which demonstrated the highest overall performance in our supporter evaluation (Section 6).

5.1 Profile Adherence

To evaluate how accurately generated utterances reflect the nine-dimensional features specified in

the input profile, we measure profile adherence.

5.1.1 Experimental Setup

Validation Pipeline and Metrics For each seeker profile, we generate a complete dialogue session with the GPT-5-mini supporter model. We then extract features from the generated seeker utterances using the LLM-based tagger and rule-based methods described in Section 3.3 and compare them with the input features using the macro F1-score.

Baselines We use a range of baselines that represent different levels of model capability and training methodologies.

- Zero-shot Base Models: GPT-4.1-mini, Llama-3-8B-Instruct (Touvron et al., 2024), Qwen-2.5-14B-Instruct (Bai et al., 2023), GPT-5 (Singh et al., 2025), and DeepSeek-V3.2 (DeepSeek-AI et al., 2025)
- SFT: We fine-tune Llama-3-8B-Instruct on our curated Reddit dialogue dataset using standard next-token prediction. Specifically, given a seeker profile and dialogue history, the model is trained to predict the next seeker utterance.
- Contrastive Learning: A variant of the SFT model that incorporates disentanglement loss to better distinguish between different feature levels. Detailed training objectives and implementation details are provided in Appendix F.

5.1.2 Results and Analysis

The experimental results, summarized in Table 1 and 13, reveal several key insights regarding the controllability of seeker simulators. Linguistic features such as **verbosity** and **dialogue length** are relatively easy to learn, with all training-based methods substantially outperforming zero-shot baselines. In contrast, psychological features like **resistance** and **self-disclosure** are much harder: zero-shot models score around 0.2 in Macro F1, standard training reaches only around 0.3, and only our MoE-based model exceeds 0.4. Overall, our model achieves the highest profile adherence across all baselines, including standard zero-shot approaches, frontier reasoning models, and other training-based methods. Feature-level accuracy and correlation scores are in Table 13 and 14.

Analysis on MoE's Adaptive Routing To identify the mechanism behind improved profile adherence, we analyze whether individual experts specialize in distinct behavioral roles.

Simulator	Mean ↑	Std ↓	Min ↑	Max ↑
GPT-4.1-mini	0.301	0.131	0.160	0.580
Llama-3-8B-Instruct	0.259	0.148	0.110	0.580
Qwen-2.5-14B-Instruct	0.284	**0.095**	0.150	0.470
GPT-5	0.319	0.216	0.150	0.840
DeepSeek-V3.2	0.431	0.218	0.180	0.910
SFT	0.515	0.160	0.360	0.760
Contrastive Learning	0.484	0.178	0.340	**0.850**
Ours	**0.549**	0.125	**0.430**	0.740

Table 1: Profile adherence results measured by Macro F1. ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Specifically, we interpret expert behavior by first assigning each dialogue to a *dominant expert*, defined as $e^* = \arg\max \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is the routing vector. We visualize the routing distributions $\boldsymbol{\alpha}$ using PCA (Figure 5), where each point is colored by its dominant expert. The clear separation of clusters indicates that the routing network partitions the representation space into distinct regions, providing a structural basis for expert-level interpretation.

We then analyze the conditional distributions $P(\text{feature} \mid e^*)$ over these assignments. These distributions are visualized as row-normalized heatmaps in Figure 6. As an example, Expert 1 shows strong concentration on high engagement and self-disclosure, along with lower resistance in the heatmap (e.g., engagement: 1.0, resistance: 0.76, self-disclosure level 4: 0.53), corresponding to a highly cooperative and open behavioral pattern.

The post-training routing analysis demonstrates that our feature-aware MoE mechanism successfully induces behaviorally meaningful specialization, aligned with distinct seeker profiles:

- *Expert 0 (Emotion-Oriented):* Specializes in seekers with high emotional distress, characterized by emotion-processing coping strategy, high resistance, and expressive (*upset/verbose*) communication.
- *Expert 1 (Collaborative & Open):* Captures highly cooperative dynamics, focusing on seekers with high engagement, low resistance, and high self-disclosure.
- *Expert 2 (Pragmatic & General):* Exhibits relatively weaker specialization, primarily handling problem-focused seekers with "average" feature levels.
- *Expert 3 (Reclusive):* Manages seekers who maintain psychological distance, utilizing

avoidant or maladaptive strategies with low engagement and minimal disclosure.

This suggests that our proposed MoE framework induces the emergent specialization of intrinsic parameter subspaces, effectively partitioning behavioral counseling regimes. Detailed figures are provided in Appendix G.

5.2 Fidelity

Fidelity focuses on whether the generated utterances reflect the authentic communicative patterns of human help-seekers. To assess this, we conduct an expert evaluation comparing our simulator with several representative baselines. All evaluations were conducted by three graduate students in clinical psychology who were trained on the evaluation instructions for each criterion.

5.2.1 Experimental Setup

Evaluation Task For each trial, experts were presented with two dialogues generated by different seeker simulators starting from the same initial problem summary. Experts select the dialogue whose seeker utterances better satisfy each evaluation criterion, or choose "Tie" if the quality is indistinguishable. Experts evaluate seeker fidelity along three dimensions: **(1) Linguistic Naturalness**, **(2) Role Authenticity**, and **(3) Psychological Plausibility**, which respectively assess language fluency, role consistency, and emotional coherence. Detailed definitions, instructions, annotator information, and inter-annotator agreement are provided in Appendix H and Table 15.

Data Sampling and Session Control We randomly sample 90 unseen seeker summaries from Reddit and fix all dialogue sessions to 10 turns to control for length effects.

Baseline Simulators We compare our model against three representative seeker simulators. To ensure that each baseline operates at its intended capacity, we adopt the original profile structures and prompting schemes specified in their respective papers. Specifically, we incorporate the seeker simulator in ESC-Judge (Madani and Srihari, 2025), a zero-shot baseline utilizing GPT-4o without additional training, and ESC-Role (Zhao et al., 2024) and Eeyore (Liu et al., 2025), both of which are fine-tuned on emotional support datasets.

Comparison	Linguistic Naturalness	Role Authenticity	Psychological Plausibility
Ours vs. Eeyore	68.9	66.7	71.1
Ours vs. ESC-Judge	68.9	72.2	62.2
Ours vs. ESC-Role	80.0	67.8	67.8

Table 2: Win rates (%) of our simulator against baseline simulators across three fidelity criteria. Win rate is computed as the proportion of wins out of total comparisons, including ties.

5.2.2 Results and Analysis

The results of the expert evaluation are summarized in Table 2. Overall, our simulator consistently outperformed all baseline models across all three dimensions, achieving an average win rate of 69.5% across all pairwise comparisons and criteria.

Specifically, our model showed a distinct advantage in linguistic naturalness over Eeyore and ESC-Role, capturing more raw communicative patterns found in real-world emotional support interactions. Furthermore, compared to ESC-Judge, our simulator demonstrated higher role authenticity, as individuals often exhibit uncertainty about their own problems rather than producing overly structured or verbose explanations. Most significantly, the superior performance in psychological plausibility validates that our simulator captures authentic patterns of emotional change, maintaining a coherent emotional trajectory across the dialogue. Detailed win/loss/tie breakdowns are provided in Table 16.

5.3 Diversity

5.3.1 Experimental Setup

To assess each simulator's coverage of diverse seeker characteristics, we evaluate the diversity of the generated utterances. Using 300 held-out profiles per simulator, we conduct dialogues with GPT-5-mini as the supporter. We compare our simulator against four baselines: ClientCAST (Wang et al., 2024a), ESC-Judge, ESC-Role, and Eeyore. Details on simulator-specific profile construction are in Appendix I. For each dialogue, seeker utterances are aggregated into a single embedding using the all-MiniLM-L6-v2 to capture the seeker's overall expressive pattern (Reimers and Gurevych, 2019).

5.3.2 Evaluation Metrics

We assess diversity using a combination of visualization-based analysis and lexical, semantic, and sentiment metrics. Detailed metric definitions and formulations are provided in Appendix J.

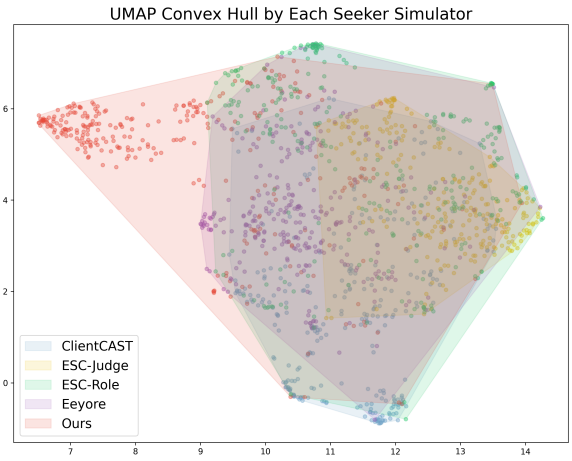


Figure 3: UMAP projection of dialogue-level seeker embeddings across simulators.

5.3.3 Results and Analysis

Figure 3 visualizes the semantic diversity of seeker utterances using UMAP projections. Our simulator exhibits substantially broader coverage compared to all baseline simulators, indicating a wider range of expressed behaviors and interaction styles. In contrast, baseline simulators tend to concentrate within narrower regions of the embedding space, suggesting limited behavioral diversity. Notably, simulators without fine-tuning (ESC-Judge and ClientCAST) occupy the most constrained regions, while SFT-based simulators (ESC-Role and Eeyore) achieve moderately broader coverage. Our MoE-based simulator achieves the largest coverage area, demonstrating that assigning dedicated experts to distinct seeker profiles effectively expands behavioral diversity.

5.3.4 Additional Quantitative Analysis

Beyond visualization, we further assess diversity using lexical, semantic, and sentiment-based metrics. Across all metrics, our simulator consistently demonstrates higher diversity than baseline models, which exhibit more repetitive language and narrower semantic and affective ranges. Detailed quantitative results are provided in Table 17.

6 Supporter Model Evaluation

In this section, we utilize our controllable seeker simulator to evaluate the performance of state-of-the-art supporter models, examining how their performance and supportive behaviors vary across diverse seeker populations. To examine the impact of seeker populations on evaluation outcomes, we evaluate the same support models using other existing seeker simulators under identical settings.

6.1 Evaluation Setup

We adopt a unified evaluation pipeline across all seeker–supporter pairs, applying the same dialogue generation and evaluation rules. All evaluations are based on multi-turn dialogues between the seeker simulator and the supporter model, followed by automated evaluation across ten distinct metrics.

Each dialogue begins with an initial seeker utterance based on the seeker's main problem. For each seeker simulator, we use 300 held-out test profiles. Dialogues terminate either via closing exchanges or, in our simulator, through explicit termination modeling using a learned `<|end_of_dialogue|>` token conditioned on the total dialogue turns level. The maximum dialogue length is set to 20 turns. A consistent supporter role prompt is used for all supporter models. Detailed evaluation settings and prompt templates are provided in Appendix K.

6.2 Evaluation Metrics

To measure the supportive capabilities of each supporter model, we adopt the expert-validated framework from Kim et al. (2025). Following the original study, all metrics are evaluated using GPT-4o-mini and are categorized into the following three groups for clarity: (1) ES Skills (Identification, Comforting, Suggestions, Experience, Informativeness), (2) General Skills (Consistency, Role-Adherence, Expression, Humanness), and (3) Overall.

6.3 Supporter Models

We evaluate a diverse range of supporter models to identify generalizable performance trends.

- Prompt-based: GPT-5-mini, Llama-3-8B-Instruct.
- Fine-tuned: Llama-3-8B-Instruct models fine-tuned on emotional support datasets, i.e., Llama-ESConv, Llama-ExTES, Llama-Psych8k, Llama-PsyInsight, and Llama-CounselChat (see Appendix L for dataset sources).

6.4 Evaluation Results

Using the unified evaluation setup in Section 6.1, we analyze how supporter performance varies across different seeker populations, highlighting the importance of population-diverse evaluation.

First, supporter model rankings vary substantially across seeker simulators. With cooperative simulators such as ESC-Judge and ESC-Role, models achieve uniformly high ES skills with small gaps, whereas under our simulator's more challenging seeker behaviors, ES scores drop sharply and model rankings change noticeably. This suggests that evaluations based on limited seeker populations can overestimate supporter robustness.

Second, evaluation metrics differ in their sensitivity to seeker behaviors. While general fluency remains relatively stable across simulators, ES skills—particularly suggestions and informativeness—exhibit larger performance drops and higher variance under our simulator. This indicates that challenging seeker behaviors primarily affect a model's ability to deliver substantive support rather than surface-level language quality (Table 18).

Third, qualitative analysis of low-scoring dialogues reveals two recurring failure patterns. With resistant seekers, models fall back on repetitive apologies instead of providing substantive empathy or constructive suggestions. With low-engagement seekers, they persist in monotonous probing rather than adapting strategies like validation or reflection. These patterns explain the observed drops in Suggestions and Informativeness.

6.5 Validity of Automatic Evaluation

To verify that the observed score drops across seeker simulators reflect genuine deficiencies in the supporter model's skills, rather than the automated judge assigning low scores in response to the seeker's negative or resistant behaviors, we conduct a human evaluation. Two graduate students in clinical psychology evaluated 60 dialogues using the same 1–5 rating scale and rubrics as the automated evaluation (details in Appendix M).

Specifically, we compute Δ_{Human} and Δ_{LLM} as the mean score differences (*Ours − ESC-Judge*) under human and automated evaluation, respectively. A negative Δ indicates that supporter models perform worse when evaluated with our simulator compared to ESC-Judge. As shown in Table 3, both Δ_{Human} and Δ_{LLM} are consistently negative across all five ES-skills, confirming that the score drops are not an artifact of automated scoring. Notably, with the exception of Experience Sharing, Δ_{Human} is larger in magnitude than Δ_{LLM} across all metrics, indicating that human experts perceive the performance gap even more strongly than the automated judge. This suggests that the score drops observed under our simulator reflect genuine supporter performance degradation, rather than systematic bias introduced by the seeker's negative behaviors.

Metric	Δ_{Human}	Δ_{LLM}	**Spearman** ρ
Identification	-1.167	-0.733	0.454
Comforting	-0.900	-0.833	0.377
Suggestions	-1.583	-1.067	0.744
Exp. Sharing	-0.417	-1.067	0.402
Informativeness	-1.917	-1.467	0.673

Table 3: Score differences (Δ = Ours − ESC-Judge) under human and automated evaluation, respectively, and their rank correlations (Spearman ρ) across ES-skills. All correlations are significant ($p < 0.05$).

6.6 Guidelines for Seeker Profile Configuration

Developers of emotional support chatbots can configure input seeker profiles according to the target seeker population of the chatbot. To illustrate how our framework can be applied in practice, we apply the same feature annotation pipeline to a real-world spoken counseling dialogue dataset (Alexander Street, 2020) and compare feature distributions between this target seeker population and our Reddit-based training data (Figure 10). These population-level differences offer a concrete basis for configuring target seeker profiles. For instance, a chatbot targeting general online users may configure profiles based on Reddit-based interactions, where seekers tend to communicate in a plain and concise style with varying levels of self-disclosure. In contrast, a chatbot targeting formal clinical psychotherapy settings may prioritize verbose utterance styles and lower self-disclosure levels (levels 1–2), reflecting the more structured and surface-level communication patterns observed in spoken counseling interactions.

7 Conclusion

We propose an evaluation framework for emotional support models based on a controllable seeker simulator, enabling assessment under diverse emotional support contexts. Our simulator demonstrates strong profile adherence, fidelity, and diversity, providing a reliable foundation for evaluation. Using this framework, we show that supporter performance varies significantly across seeker populations, revealing that conventional evaluations can mask meaningful weaknesses and overestimate real-world support effectiveness.

Limitations

Our work demonstrates that a controllable and population-diverse seeker simulator enables systematic evaluation of emotional support models across a wide range of interaction settings. However, our framework evaluates emotional support within a single dialogue, and therefore does not capture the cumulative nature of emotional support effects, which often unfold across repeated interactions in real-world settings.

In addition, we represent seekers using profiles defined by combinations of behavioral features to enable controlled comparison. While this abstraction robustly captures diverse seeker characteristics, some features may naturally evolve over the course of an interaction. Although such variation may be partially reflected in the training data, dynamically changing features are treated as fixed during evaluation, and finer-grained control over temporal feature transitions remains an open direction.

Finally, our evaluation prioritizes supporting skills and overall conversational quality, using ES skills and general skills to assess supporter performance under diverse seeker populations. While this reveals meaningful performance differences across populations, it does not directly measure longer-term outcomes such as sustained emotional recovery, well-being, or behavioral change. Incorporating complementary evaluation settings and outcome-oriented metrics remains a promising direction for future extensions of this framework.

Acknowledgments

This work was supported by the Creative-Pioneering Researchers Program through Seoul National University and by the National Research Foundation of Korea (NRF) grants (RS-2024-00333484 and RS-2024-00414981) funded by the Korean government (MSIT).

We used AI assistants to proofread the writing and to help with coding.

References

Alexander Street. 2020. Counseling and psychotherapy transcripts series ii. Alexander Street, a ProQuest Company. Commercial counseling dialogue corpus.

Irwin Altman and Dalmas A. Taylor. 1973. *Social Penetration: The Development of Interpersonal Relationships*. Holt, Rinehart & Winston, New York, NY.

Jinze Bai et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Nicolas Bertagnolli. 2020. Counselchat: Bootstrapping high-quality therapy data. https://github.com/nbertagnolli/counsel-chat. Accessed: 2025-12-20.

Sharon S. Brehm and Jack W. Brehm. 1981. *Psychological Reactance: A Theory of Freedom and Control*. Academic Press.

Keqi Chen, Zekai Sun, Yuhua Wen, Huijun Lian, Yingming Gao, and Ya Li. 2025. Psy-insight: Explainable multi-turn bilingual dataset for mental health counseling. *Preprint*, arXiv:2503.03607.

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.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Erhang Li, Fangqi Zhou, Fangyun Lin, Fucong Dai, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoran Wei, Haowei Zhang, Haowen Luo, Haozhe Ji, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, Jialiang Huang, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jingchang Chen, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexin Huang, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Liang Zhao, Liangsheng Yin, Lihua Guo, Lingxiao Luo, Linwang Ma, Litong Wang, Liyue Zhang, M. S. Di, M. Y Xu, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Panpan Huang, Peixin Cong, Peiyi Wang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Songyang Zhou, Tao Ni, Tao Yun, Tian Pei, Tian Ye, Tianyuan Yue, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjun Gao, Wentao Zhang, Xi Gao, Xiangwen Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyuan Li, Xu Chen, Xuecheng Su, Xuehai Pan, Xuheng Lin, Xuwei Fu, Y. Q. Wang, Yang Zhang, Yanhong Xu, Yanru Ma,

Yao Li, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yiliang Xiong, Ying He, Ying Zhou, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuduan Wang, Yue Gong, Yuhan Wu, Yuheng Zou, Yukun Li, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehua Zhao, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhiyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Zizheng Pan, Zongqing Yao, Bei Feng, Hui Li, J. L. Cai, Jiaqi Ni, Lei Xu, Meng Li, Ning Tian, R. J. Chen, R. L. Jin, S. S. Li, Shuang Zhou, Tianyu Sun, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xinnan Song, Xinyi Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Dongjie Ji, Jian Liang, Jianzhong Guo, Jin Chen, Leyi Xia, Miaojun Wang, Mingming Li, Peng Zhang, Ruyi Chen, Shangmian Sun, Shaoqing Wu, Shengfeng Ye, T. Wang, W. L. Xiao, Wei An, Xianzu Wang, Xiaowen Sun, Xiaoxiang Wang, Ying Tang, Yukun Zha, Zekai Zhang, Zhe Ju, Zhen Zhang, and Zihua Qu. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Norman S. Endler and James D. A. Parker. 1994. Assessment of multidimensional coping: Task, emotion, and avoidance strategies. *Psychological Assessment*, 6(1):50–60.

Emma Holdsworth, Erica Bowen, Sarah Brown, and Douglas Howat. 2014. Client engagement in psychotherapeutic treatment and associations with client characteristics, therapist characteristics, and treatment factors. *Clinical Psychology Review*, 34(5):428–450.

Juhee Kim, Chunghu Mok, Jisun Lee, Hyang Sook Kim, and Yohan Jo. 2025. Dialogue systems for emotional support via value reinforcement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28733–28766, Vienna, Austria. Association for Computational Linguistics.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada. Association for Computational Linguistics.

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

Siyang Liu, Bianca Brie, Wenda Li, Laura Biester, Andrew Lee, James Pennebaker, and Rada Mihalcea. 2025. Eeyore: Realistic depression simulation via expert-in-the-loop supervised and preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13750–13770, Vienna, Austria. Association for Computational Linguistics.

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.

Navid Madani and Rohini Srihari. 2025. ESC-judge: A framework for comparing emotional support conversational agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16059–16076, Suzhou, China. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. Https://arxiv.org/abs/2303.08774.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *Preprint*, arXiv:2408.15787.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Sashank Santhanam and Samira Shaikh. 2019. Emotional neural language generation grounded in situational contexts. In *Proceedings of the 4th Workshop on Computational Creativity in Language Generation*, pages 22–27, Tokyo, Japan. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao,

Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Hugo Touvron et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024a. Towards a client-centered assessment of llm therapists by client simulation. *Preprint*, arXiv:2406.12266.

Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. 2025. AnnaAgent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23221–23235, Vienna, Austria. Association for Computational Linguistics.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024b. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Neslihan Yaman. 2021. Working with resistance in therapy: A theoretical evaluation. *IBAD Journal of Social Sciences*, 9:481–495.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling Kit, Jenny Giam Xiuhui, John Pinto, and Ee-Peng Lim. 2025. Consistent client simulation for motivational interviewing-based counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20959–20998, 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.

Xi Zheng, Zhuoyang Li, Xinning Gui, and Yuhan Luo. 2025. Customizing emotional support: How do individuals construct and interact with llm-powered chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–20. ACM.

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.

Yuhang Zhou, Zihua Zhao, Siyuan Du, Haolin Li, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. 2024. Exploring training on heterogeneous data with mixture of low-rank adapters. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

A Data Preprocessing Details

In addition to the general data quality criteria described in the main text, we applied several concrete preprocessing procedures to the dataset.

We retained only dialogues consisting of at least five turns to ensure sufficient interactional context for modeling seeker behavior. Topic-based filtering was performed using GPT-4o-mini to exclude threads unrelated to mental health, such as requests for investment advice or restaurant recommendations. To ensure stable model training, dialogues containing non-English utterances were removed. We also removed emojis and other non-standard symbolic characters to reduce noise and variability in textual representations. Finally, a minimum upvote threshold was applied to filter out low-quality or low-engagement content.

These steps were applied uniformly across the dataset prior to training.

B Feature Category Definitions

To enable fine-grained control over diverse seeker behaviors, we define a set of fine-grained categories for each of the nine features in our seeker profile taxonomy. These categories are designed to capture subtle yet practically meaningful variations in how seekers express emotions, engage with counselors, and respond throughout emotional support dialogues. The category-level definitions for each feature are provided below.

B.1 Coping Strategy

- **problem_focused**: Engages in concrete actions or cognitive efforts (e.g., planning, evaluating options, or using external resources) to address or change the stressor.
- **emotion_processing**: Focuses on expressing, exploring, or emotionally processing feelings rather than directly solving the problem.
- **avoidant**: Deliberately avoids thinking about, confronting, or engaging with the problem or its associated emotions.
- **maladaptive_behavior**: Responds to stress through currently ongoing self-destructive or harmful behaviors.

B.2 Utterance Style

While the original framework suggests a broader range of styles, we focus on the three most prevalent styles identified in our Reddit corpus—*plain*, *upset*, and *verbose*—excluding those with negligible frequency in online support settings.

- **plain**: Communicates thoughts and answers in a direct, concise, and emotionally neutral manner.
- **upset**: Expresses strong negative emotions such as anger or frustration in a confrontational or resistant tone.
- **verbose**: Produces excessively long and detailed utterances that disrupt conversational flow.

B.3 Resistance Level

- **high**: Displays persistent and pervasive resistance across most topics and counselor interventions throughout the dialogue.
- **medium**: Shows situational resistance triggered by specific topics or interventions, alternating with periods of cooperation.
- **low**: Is generally cooperative, with resistance absent or limited to brief and minor instances.

B.4 Engagement Level

- **high**: Actively participates as an equal partner by reflecting, expressing emotions, and driving the conversation forward.
- **medium**: Participates inconsistently, with engagement limited to specific dimensions or topics.
- **low**: Demonstrates minimal willingness to engage, providing short, dismissive, or disengaged responses.

B.5 Self-Disclosure Level

- **1 (orientation)**: Shares only surface-level, socially normative information without emotional content or personal meaning.
- **2 (exploratory affective exchange)**: Shares personal facts or light emotions without deep vulnerability.
- **3 (affective exchange)**: Reveals emotionally meaningful experiences, fears, or vulnerabilities tied to personal significance.
- **4 (stable exchange)**: Expresses identity-level beliefs, core assumptions, or existential conclusions that generalize beyond specific events.

B.6 Client Reaction Proportions

- **positive**: Responds to the counselor with openness, cooperation, or constructive engagement.
- **neutral**: Responds without clear acceptance or rejection, remaining vague or minimally responsive.
- **negative**: Pushes back against, rejects, or deflects the counselor's message in a resistant or defensive manner.

B.7 Profanity Flag

- **true**: Contains profanity or explicit offensive language.
- **false**: Does not contain profanity or explicit offensive language.

B.8 Verbosity Level

- **very_short**: Fewer than 15 tokens.
- **short**: 15–29 tokens.
- **medium**: 30–59 tokens.
- **long**: 60–99 tokens.
- **very_long**: 100 tokens or more.

B.9 Total Dialogue Turns Level

- **short**: 4–5 turns.
- **medium**: 6–8 turns.
- **long**: 9 turns or more.

C LLM-based Psychological Feature Tagging

To capture both stable interaction patterns and turn-level variations, we perform feature tagging at two levels—dialogue level and utterance level—and subsequently aggregate the annotations into dialogue-level representations used for model training and evaluation.

- **Dialogue-level Features:** *Main coping strategy*, *utterance style*, *resistance level*, and *engagement level* are annotated once per dialogue. Since these features reflect the seeker's dominant interaction patterns, the LLM processes the entire dialogue history to infer a single representative label. See Table 10 for detailed prompt specifications.
- **Utterance-level Features:** In contrast, *self-disclosure level* and *seeker reaction proportions* may vary across turns depending on the conversational context including preceding supporter responses. Accordingly, these features are tagged at each turn based on supporter–seeker utterance pairs to capture such dynamic shifts. To construct final dialogue-level profile, *self-disclosure level* scores are averaged and rounded, while *seeker reaction proportions* is summarized as a distribution of positive, neutral, and negative responses. The common system instruction is provided in Figure 4, and detailed prompt specifications are listed in Table 11.

Feature	IAA (Percent Agr.)	Human–LLM (Accuracy)
Main coping strategy	0.442	0.729
Utterance style	0.528	0.852
Resistance level	0.694	0.902
Engagement level	0.597	0.854
Self-disclosure level	0.564	0.827
Seeker reaction	–	0.860

Table 4: Feature-level inter-annotator agreement and human–LLM alignment.

Common system instruction

```
You are an expert evaluator. You will be
provided with the client's main problem for
context, followed by a multi-turn dialogue
between a counselor and that client.
```

Figure 4: Common system instruction for feature tagging.

D Human Validation for LLM-based Tagging

To enhance annotation consistency, evaluators were randomly assigned to three specialized groups based on feature characteristics:

- **Group A**: Evaluates psychological and expressive characteristics (*main coping strategy*, *utterance style*).
- **Group B**: Evaluates interactional attitudes (*resistance level*, *engagement level*).
- **Group C**: Evaluates turn-level dynamics (*self-disclosure level*, *seeker reaction*).

Feature-level inter-annotator agreement (IAA) and human–LLM alignment results are reported in Table 4.

E Rule-based Linguistic Feature Extraction

E.1 Linguistic Feature Extraction Rules

Linguistic features are extracted using deterministic, rule-based procedures. The *verbosity level* is computed as the average token count of seeker utterances, excluding the initial turn corresponding to the original Reddit post, and discretized into a five-level scale based on the empirical distribution of the dataset. The *user profanity flag* is extracted using the `profanity-check` library and assigned as a binary value, where the flag is set to 1 if any seeker utterance exceeds a predefined probability threshold for profanity, and 0 otherwise. The *total dialogue turns level* is derived from the total number of turns in the dialogue and categorized into three tiers—Short, Medium, and Long—according to the dataset's distribution.

These structural features provide objective signals that complement the psychological annotations by reflecting surface-level behavior and overall dialogue complexity.

F Implementation Details for Contrastive Learning

This section provides the mathematical formulation of the proposed **Disentanglement Loss (L_D)** and the technical specifications for pseudo-feature generation in contrastive learning baseline.

Training Objective

The model is trained by combining the standard language modeling loss (L_{LM}) with our proposed Disentanglement Loss (L_D). The total training objective is defined as:

$$L_{total} = L_{LM} + \lambda_D L_D$$

L_{LM} denotes the standard cross-entropy loss on the original dataset. L_D is designed to ensure that the model generates responses that are clearly distinguishable based on the provided seeker features. The formulation of L_D is as follows:

$$L_D = -\log \frac{P(y \mid x_o)}{P(y \mid x_o) + \sum_{i=1}^{N} P(y \mid x_{p,i})}$$

where:

- x_o: The original input.
- $x_{p,i}$: The i-th pseudo-input where features are intentionally modified.
- y: The ground-truth next client utterance from the original data.
- N: The number of pseudo-samples generated per anchor (set to $N = 3$ in this study).

By maximizing the log-likelihood of the ground-truth response under the original features relative to the modified features, L_D forces the model to be highly sensitive to the specific feature descriptions in the system prompt.

Pseudo-feature Generation

To construct the contrastive set for L_D, we strategically manipulate the `seeker_features` block within the system prompt. For each sample, three pseudo-variants are generated using the following priority rules:

- **Extreme Flip**: Numerical or ordinal features such as *engagement level*, *resistance level*, and *self-disclosure level* are flipped to their opposite extremes (e.g., *high* $\leftrightarrow$ *low*).
- **Categorical Shift**: Categorical features like *main coping strategy* or *utterance style* are replaced with a randomly selected alternative category.
- **Medium Flip**: Neutral values (e.g., *medium*) are randomly reassigned to extreme values to enhance the discriminative signal.

Training Configuration

The model was trained using the hyperparameters summarized in Table 5.

Table 5: Training Configuration for CL

Category	Configuration
Base Model	Meta-Llama-3-8B-Instruct
LoRA r / α	16 / 16
LoRA Target	All Linear Layers
LoRA Dropout	0.05
Learning Rate	2×10^{-5}
Epochs	1
Batch Size	16 (4 per device $\times$ 4 GA)
Max Length	2048
λ_D	1.0

G Routing Analysis in MoE

Figure 5 visualizes the routing space via PCA. Each point corresponds to a dialogue sample and is colored by its dominant expert $e^* = \arg\max \boldsymbol{\alpha}$. Distinct clustering by dominant expert suggests that the routing network learns separable regions in $\boldsymbol{\alpha}$

space, reinforcing the observed feature-conditioned specialization. This structural separation indicates that routing decisions are organized in a meaningful latent space, where each expert occupies a consistent region, supporting expert-level interpretation. Figure 6 presents conditional distributions $P(\text{feature} \mid e^*)$ of behavioral features given the dominant expert. Each row is normalized to sum to 1. High values in a row indicate that a particular feature level is strongly associated with that expert. These patterns support the claim that experts capture distinct behavioral regimes rather than surface-level variations.

These distributions reveal distinct behavioral tendencies for each expert. For example, Expert 1 shows strong concentration on high engagement (1.0), low resistance (0.76), and high self-disclosure (level 4: 0.53), whereas Expert 3 exhibits relatively higher proportions of low engagement and lower self-disclosure levels (1–2). Such patterns demonstrate that expert routing aligns with coherent behavioral profiles rather than superficial variation.

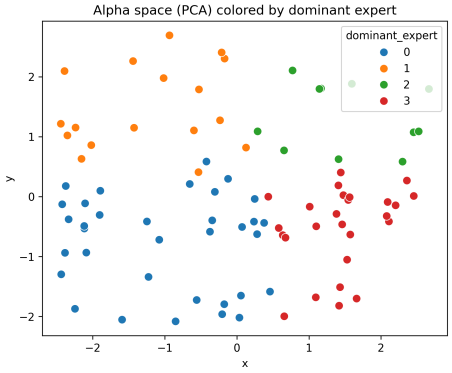


Figure 5: PCA projection of routing distributions (α), colored by dominant expert. The separation indicates that expert routing occupies distinct regions in routing space.

Despite this expert-based structure, the MoE framework incurs minimal overhead. The routing network adds only 15,881 parameters ($\sim$0.0003% of total model size), requiring approximately 2 hours and 40 minutes of additional training on a single H100 GPU. Moreover, MoE provides targeted improvements on features that SFT largely fails to control, such as self-disclosure level (0.40 $\rightarrow$ 0.45) and resistance level (0.37 $\rightarrow$ 0.43), effectively expanding the range of reliably controllable behavioral dimensions.

H Expert Evaluations on Fidelity

H.1 Metric Definitions

- Linguistic Naturalness: Does the seeker speak in a fluid, human-like manner without mechanical repetitions or awkward phrasing?
- Role Authenticity: Does the seeker maintain the role and behavioral traits consistent with a person seeking emotional support?
- Psychological Plausibility: Are the seeker's emotional transitions and reactions psychologically coherent throughout the conversation?

H.2 Expert Evaluation Settings

For expert evaluation, evaluators were first instructed on the criteria-specific guidelines, and then performed the assessment using the interface in Figure 9 based on the instructions in Table 15.

Annotator Details The evaluation was conducted by three graduate students specializing in clinical psychology (two males and one female). Participation was voluntary, and each annotator was compensated at approximately $40 per hour in accordance with standard research participation guidelines.

Inter-Annotator Agreement We report inter-annotator agreement to assess the reliability of our expert evaluations. Annotators were asked to choose among "prefer A", "prefer B", and "prefer both". Given the ordinal nature of the scale (A = 0, Both = 0.5, B = 1), we compute mean absolute ordinal deviation to capture partial agreement rather than treating all disagreements as equally distant. The resulting scores were 0.341, 0.315, and 0.322 for Linguistic Naturalness, Role Authenticity, and Psychological Plausibility, respectively, indicating that annotators rarely showed completely opposite preferences. Pairwise agreement and full agreement rates are reported in Table 6.

Metric	1–2	1–3	2–3	All
Linguistic Naturalness	0.811	0.433	0.411	0.367
Role Authenticity	0.744	0.411	0.389	0.311
Psychological Plausibility	0.800	0.300	0.311	0.267

Table 6: Pairwise and full inter-annotator agreement rates. Columns 1–2, 1–3, 2–3 denote pairwise agreement between annotators, and All denotes the rate of full agreement across all three annotators.

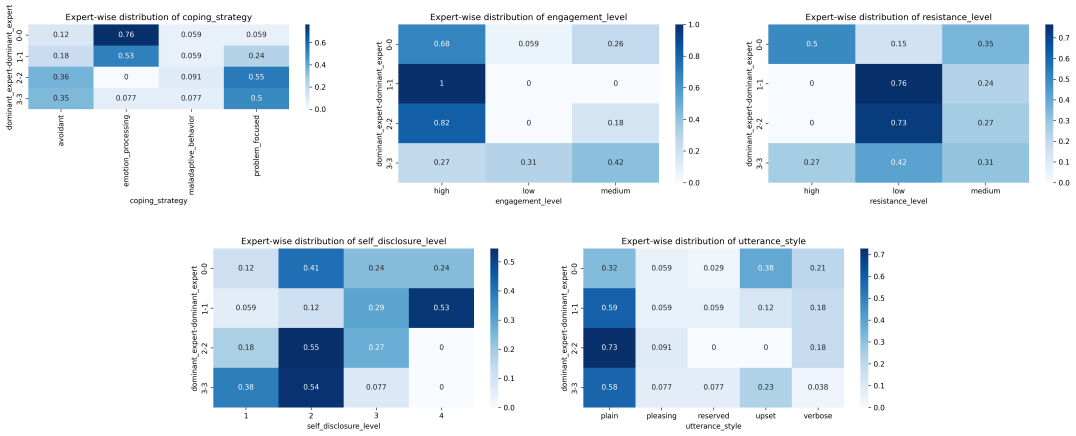


Figure 6: Each heatmap row is normalized to sum to 1, highlighting expert preference under each feature label.

I Profile construction for each Seeker Simulator

To ensure fair comparison and preserve the originally reported performance of each seeker simulator, we adopt the original profile structures and prompting schemes proposed in their respective papers without modification. We construct 300 held-out Reddit-derived profiles for each simulator following its originally proposed profile construction procedure. For ClientCAST, we instead use its originally proposed datasets directly, as the simulator requires reference dialogues to be provided during generation. In addition, when constructing simulator-specific profiles, we aim to cover a broad range of seeker characteristics within each simulator's representational capacity by assigning diverse feature labels where applicable.

ESC-Judge For ESC-Judge, we construct seeker profiles following the original profile generation framework proposed in the paper. Each profile is generated by sampling an ongoing challenge from a predefined stressor set and composing a structured persona that includes demographic attributes, family context, and occupational background. Additional contextual information, such as past life experiences and behavioral traits is sampled from predefined spaces and integrated into a unified profile card.

ESC-Eval For ESC-Eval, we directly use the publicly released test profiles provided by the original work. To maintain diversity while matching the evaluation scale of other simulators, we randomly sample 300 profiles from the released test set, preserving the original distribution of profile attributes.

Eeyore For Eeyore, the number of publicly available test profiles is smaller than the required scale. To address this, we construct additional profiles following the procedure described in the original paper. Specifically, we extract each seeker's main problem from the training dataset using the same extraction method employed in our Reddit data preprocessing pipeline and assign it to the client situation category. We then assign categorical labels to promote diversity across feature dimensions, such as varying symptom severity, balancing resistance toward support, and diversifying persona attributes, while maintaining balanced coverage across feature values.

ClientCAST For ClientCAST, we reproduce the simulator following the methodology and codebase described in the original paper. Seeker profiles are constructed from two counseling datasets: the High-Low Quality Counseling dataset (Pérez-Rosas et al., 2019) and AnnoMI (Wu et al., 2022), from which we extract 213 high- and 87 low-quality sessions as specified in the original work. As ClientCAST requires reference dialogues during simulation, we directly use the original counseling sessions from these datasets. Each seeker profile consists of three components: (1) Problems & Reasons for Visiting, (2) Displayed Symptoms from a predefined set of 61 client symptoms, and (3) Apparent Traits, assigned to promote diversity across seeker characteristics.

J Diversity Metrics

We employ embedding-space visualizations together with lexical, semantic, and sentiment-based metrics to capture the diversity of the simulators' outputs.

Seeker Simulator	TSNE Hull Area ↑	PCA Hull Area ↑
ClientCAST	2954.562	0.312
ESC-Judge	1878.463	0.334
ESC-Role	3093.516	0.378
Eeyore	3057.489	0.441
Ours	**3378.139**	**0.659**

Table 7: Convex hull areas computed on t-SNE and PCA projections of dialogue-level seeker embeddings, reflecting the spatial coverage of each simulator.

J.1 Visualization Metrics

UMAP-based Visualization Metric We visualize seeker diversity by projecting dialogue-level embeddings into a two-dimensional space using UMAP (Uniform Manifold Approximation and Projection). Each dialogue is represented by a single embedding obtained by aggregating all seeker utterances and encoding them with a sentence embedding model. UMAP is applied to the combined embeddings from all seeker simulators using cosine distance, preserving local neighborhood structure while maintaining coarse global relationships in the low-dimensional space.

Convex Hull-based Coverage Metrics To complement the qualitative patterns observed in the UMAP visualizations, we additionally quantify the spatial coverage of each seeker simulator using convex hull area metrics. Specifically, we compute the area of the convex hull enclosing dialogue-level embeddings projected onto two-dimensional spaces obtained via t-SNE and PCA. These metrics provide a simple geometric estimate of how broadly each simulator's generated dialogues are distributed in the embedding space, serving as a quantitative proxy for the visual spread observed in the projections.

Table 7 reports the convex hull areas for each simulator. Consistent with the UMAP visualizations, our simulator exhibits the largest hull area under both t-SNE and PCA projections, indicating broader coverage of the embedding space compared to baseline simulators. In contrast, baseline simulators occupy more confined regions, reflecting more limited diversity in generated seeker expressions.

J.2 Quantitative Metrics

In addition to the visualization-based analysis described above, we employ a set of quantitative diversity metrics to measure lexical variation, semantic dispersion, and sentiment distribution in a complementary and reproducible manner. These metrics provide numerical estimates of diversity that support the qualitative patterns observed in the embedding-space visualizations.

Metric Definitions

- **Lexical Diversity** We measure lexical diversity using Distinct-n, Self-BLEU, and Token Repetition Mean, computed over dialogue-level seeker texts. Distinct-n is calculated as the ratio of unique n-grams to the total number of n-grams across all dialogues, capturing surface-level vocabulary variety. Self-BLEU is computed by treating each dialogue as a hypothesis and the remaining dialogues as references, and averaging BLEU scores across dialogues to quantify intra-set redundancy. Token Repetition Mean is calculated as the average proportion of repeated tokens within each dialogue, reflecting repetitive word usage.
- **Semantic Diversity** To assess semantic diversity, we compute Average Pairwise Distance (APD) and Average Cosine Similarity (ACS) over dialogue-level embedding representations. Each dialogue is embedded by encoding the concatenation of all seeker utterances using a sentence embedding model. APD is computed as the mean cosine distance over all unique pairs of dialogue embeddings, measuring the overall semantic dispersion of generated dialogues. ACS is computed as the mean cosine similarity over the same embedding pairs, capturing the degree of semantic redundancy. Higher APD (and correspondingly lower ACS) indicates greater semantic diversity among dialogue representations.
- **Sentiment Diversity** We evaluate sentiment diversity using VADER compound sentiment scores computed for each dialogue. For each simulator, we report the mean and variance of compound scores across dialogues. The mean reflects the overall sentiment tendency, while the variance captures the spread of expressed emotional polarity across generated seeker utterances.

Results and Analysis Across all lexical and semantic diversity metrics, our simulator consistently achieves the highest diversity scores.

Specifically, it attains the strongest performance on lexical diversity measures including Distinct-

2, Self-BLEU, and Token Repetition Mean, indicating richer and less repetitive vocabulary usage. Similarly, on semantic diversity metrics—Average Pairwise Distance (APD) and Average Cosine Similarity (ACS)—our simulator exhibits the largest semantic dispersion and the lowest redundancy among dialogue representations.

We further analyze affective diversity using VADER compound scores. Our simulator yields a sentiment mean of -0.05 with a variance of 0.70, reflecting a wide distribution of emotional states spanning from negative to positive affect. In contrast, baseline simulators exhibit sentiment means above 0.7 with variances below 0.5, suggesting a strong bias toward uniformly positive expressions and limited emotional variability. These results indicate that existing simulators tend to generate overly optimistic seeker behaviors, whereas our model more faithfully captures the emotional heterogeneity observed in real-world emotional support seekers, including both negative and positive affective states.

K Evaluation Settings for Supporter Model Evaluation

Evaluation Settings To ensure comparable initial conditions across evaluations, dialogues are initialized using seeker utterances derived from the seeker's main problem. This design choice standardizes the starting context and prevents variations in supporter performance caused by inconsistent dialogue entry points.

All seeker simulators are evaluated using held-out profiles that preserve the original profile structures and prompting schemes specified in prior work. This strategy maintains the intended operating conditions of each simulator while enabling controlled comparison across different seeker populations.

Dialogue termination is implemented differently depending on the simulator. In standard settings, conversations naturally conclude through mutual closing exchanges, such as expressions of gratitude or farewell. In contrast, our seeker simulator explicitly models dialogue termination by conditioning generation on the total dialogue turns level feature and emitting a dedicated `<|end_of_dialogue|>` token. This mechanism is learned from training data and allows the simulator to regulate dialogue length in a principled and interpretable manner.

The maximum dialogue length constraint is informed by empirical statistics of the underlying data, where seeker–supporter interactions exhibit an average length of approximately 17 turns. This observation motivates the dialogue length cap used in evaluation.

Prompts for Supporter Models To ensure fairness across supporter models, a unified supporter role prompt is applied throughout all evaluations. The full prompt template is provided in Figure 7.

Supporter role prompt

```
You are a supportive and empathetic
assistant for an emotional support context.
Respond in a calm and non-judgmental manner,
acknowledging and validating the user's
emotions.
Ask gentle clarifying questions only when
helpful, and offer practical coping
suggestions when appropriate.
Do not diagnose, prescribe, or claim
professional authority.
Keep responses concise and warm (3–6
sentences).
```

Figure 7: Supporter role prompt template for the emotional support simulator.

L Datasets Used for Fine-tuning Supporter Models

The supporter models were fine-tuned on publicly available emotional support datasets, including ESConv (Liu et al., 2021), ExTES (Zheng et al., 2024), Psych8k (Liu et al., 2023), PsyInsight (Chen et al., 2025), and CounselChat (Bertagnolli, 2020).

M Human Evaluation Details

Experimental Setup We selected Ours and ESC-Judge as the two seeker simulators, as they produced the largest performance gaps in supporter evaluation. For supporter models, we included GPT-5-mini and Llama-ESConv, representing prompt-based and fine-tuned supporters, respectively. From the dialogues generated between these seeker simulators and supporter models, we sampled 60 dialogues ensuring a uniform distribution of score discrepancies, covering the full spectrum from cases with similar scores to those with substantial differences. Beyond the Δ-based analysis presented in Section 6.5, we additionally report Human–LLM score correlations and inter-annotator agreement to verify the reliability of the

automated evaluation and the human evaluation itself, respectively.

Human–LLM Raw Score Correlations To assess whether the automated scoring aligns with human judgments at the absolute score level, we computed Pearson and Spearman correlations between human and automated scores. The resulting correlations were moderate to strong across all ES-skills (all $p < 0.001$), indicating that the automated evaluation reliably reflects human assessments. (Table 8)

Metric	**Pearson** r	**Spearman** ρ
Identification	0.650	0.628
Comforting	0.589	0.546
Suggestions	0.730	0.761
Exp. Sharing	0.636	0.541
Informativeness	0.807	0.794

Table 8: Pearson and Spearman correlations between human and automated evaluation scores.

Inter-Annotator Agreement To verify the reliability of the human evaluation itself, we assessed inter-annotator agreement using quadratic weighted Cohen's κ, which penalizes larger disagreements more heavily. Agreement ranged from fair to strong across metrics, indicating generally reliable human evaluation. (Table 9)

Metric	**Quadratic** κ
Identification	0.595
Comforting	0.345
Suggestions	0.658
Exp. Sharing	0.821
Informativeness	0.774

Table 9: Inter-annotator agreement measured by quadratic weighted Cohen's κ.

Feature	LLM Tagging Prompt
Main Coping Strategy	**Task.** Evaluate the client's main coping strategy, defined as the single most dominant strategy the client uses across the entire dialogue to manage stress, based only on the client's utterances. **Instructions.** (1) Read the entire dialogue. Read the client's main problem for context only. (2) Base the evaluation only on the client's utterances. (3) Assign a single coping strategy label that best represents the client's dominant and consistent approach. **Labels.** *problem_focused*: concrete actions or cognitive efforts to address the stressor; *emotion_processing*: emotional expression or processing without problem solving; *avoidant*: deliberate avoidance of the problem or emotions; *maladaptive_behavior*: current self-destructive coping behavior.
Engagement Level	**Task.** Evaluate the client's overall engagement level throughout the entire dialogue, reflecting how active, serious, and honest the client is in the interaction. **Instructions.** (1) Read the entire dialogue. Read the client's main problem for context only. (2) Base the evaluation only on the client's utterances. (3) Assign a single engagement level representing the client's overall mode of participation. **Labels.** *high*: active and reflective participation; *medium*: inconsistent or biased engagement; *low*: minimal or disengaged participation.
Resistance Level	**Task.** Evaluate the overall intensity and consistency of resistance behaviors exhibited by the client across the entire dialogue. **Instructions.** (1) Read the entire dialogue. Read the client's main problem for context only. (2) Base the evaluation only on the client's utterances. (3) Assign a single resistance level reflecting the client's dominant behavioral pattern. **Labels.** *high*: frequent and pervasive resistance; *medium*: situational resistance triggered by specific topics or interventions; *low*: generally cooperative with little or no resistance.
Utterance Style	**Task.** Evaluate the client's dominant utterance style used throughout the dialogue, focusing on how the client speaks rather than what is said. **Instructions.** (1) Read the entire dialogue. Read the client's main problem for context only. (2) Base the evaluation only on the client's utterances. (3) Assign a single utterance style that best characterizes the client's speaking pattern. **Labels.** *plain*: direct, clear, and to-the-point communication without excessive emotion or avoidance; *upset*: expression of frustration, anger, or strong resistance in a confrontational or dismissive tone; *verbose*: excessively long and detailed utterances that disrupt conversational flow; *reserved*: minimal, vague, or evasive responses requiring repeated prompting; *tangent*: deviation from the main topic into unrelated content; *pleasing*: a persistent pattern of excessive agreement or problem minimization toward the counselor.

Table 10: LLM prompts used for dialogue-level feature tagging.

Feature	LLM Tagging Prompt
Seeker Reaction	**Task.** Evaluate how the seeker reacts to the counselor's immediately preceding utterance, focusing only on the seeker's response rather than the seeker's general emotional state or situation. **Instructions.** (1) Read the client's main problem for context only. (2) Read the counselor utterance and the immediately following seeker utterance. (3) Base the evaluation only on the seeker's utterance. (4) Assign a single reaction label that best represents the seeker's response to the counselor. **Labels.** *positive*: openness, cooperation, acknowledgment, agreement, appreciation, or constructive engagement with the counselor's message; *neutral*: vague, minimal, or non-committal responses without clear acceptance or rejection; *negative*: rejection, resistance, defensiveness, deflection, or pushback toward the counselor's message.
Self-Disclosure Level	**Task.** Evaluate the depth of personal information or emotional vulnerability revealed in the seeker's utterance in response to the counselor's immediately preceding message. **Instructions.** (1) Read the client's main problem for context only. (2) Read the counselor utterance and the immediately following seeker utterance. (3) Base the evaluation only on the seeker's utterance. (4) Assign a single self-disclosure level reflecting the depth of revealed information. **Labels.** *1 (orientation)*: surface-level, socially normative information without emotional content or personal meaning; *2 (exploratory affective exchange)*: personal facts or light emotions without deep vulnerability; *3 (affective exchange)*: emotionally meaningful experiences, fears, or vulnerabilities with personal significance; *4 (stable exchange)*: identity-level beliefs, core assumptions, or existential conclusions that generalize beyond specific events.

Table 11: LLM prompts used for utterance-level feature tagging.

<seeker_features>
coping_strategy: emotion_processing
resistance_level: high
engagement_level: high
self_disclosure_level: 2
seeker_reaction: {"positive_ratio": 0.0, "neutral_ratio": 0.0, "negative_ratio": 1.0}
utterance_style: upset
verbosity_level: medium (30–59 tokens)
user_profanity_flag: True
total_dialogue_turns_level: long ($\geq$ 9 turns)

<seeker's_main_problem>
The seeker is struggling with feelings of failure after not being accepted into medical school, despite having strong academic credentials. This obsession with medicine overshadows their ability to focus on job searching or other career paths, leading to persistent emotional distress and self-doubt.

Figure 8: An example of seeker profile.

System Prompt	You are a summarization assistant.
User Prompt	You are an evaluator. You will be provided with a first-person Reddit post written by a seeker. Your task is to write a concise summary that captures the seeker's situation and main problem. Please follow these steps: 1. Read the post carefully and extract the context (what happened) and the core concern or emotional struggle. 2. Write the summary in the third person (do not use "I", "my", etc). 3. The summary must begin with "problem:" 4. Keep the summary to 1–2 sentences. 5. Do not include any information that is not clearly stated in the post. 6. Do not include any names or proper nouns. You may mention age or gender only when it is explicitly mentioned in the post, and never guess or infer it. [Post] {OP text} [Output format] problem: [summary in third person, 1–2 sentences, capturing the situation and the seeker's emotional or psychological struggle]

Table 12: Prompt template for extracting seeker's main problem from first-person Reddit posts.

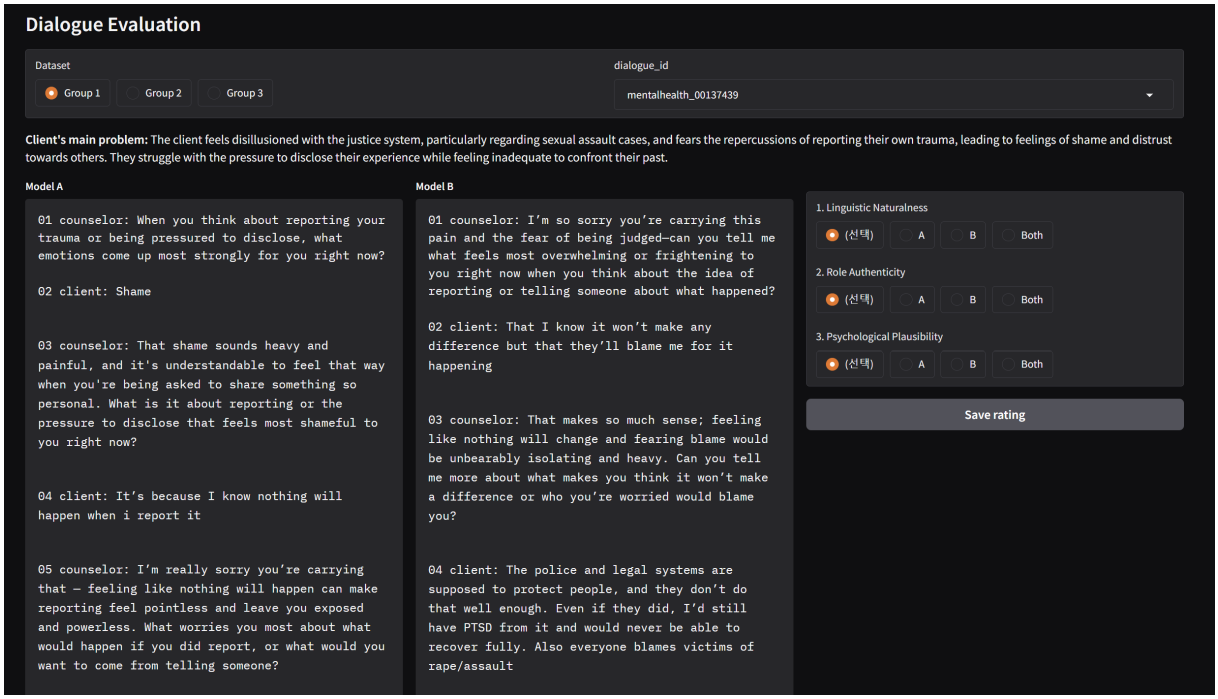


Figure 9: User interface for expert evaluation.

Model	Coping	Utterance	Engagement	Verbosity	Self-Disclosure	Resistance	Profanity	Turns
GPT-4.1-mini	0.33	0.36	0.27	0.28	0.19	0.24	0.58	0.16
Llama-3-8B-Instruct	0.29	0.24	0.27	0.11	0.14	0.29	0.58	0.15
Qwen-2.5-14B-Instruct	0.25	0.32	0.29	0.33	0.22	0.24	0.47	0.15
GPT-5	0.23	0.33	0.27	0.23	0.25	0.25	0.84	0.15
DeepSeek-V3.2	0.39	0.40	0.41	0.52	0.27	0.37	**0.91**	0.18
SFT	0.38	0.36	0.53	0.66	0.40	0.37	0.66	0.56
Contrastive Learning	0.30	0.32	0.39	0.54	0.32	0.35	0.64	**0.83**
Ours	**0.44**	**0.47**	**0.58**	**0.69**	**0.45**	**0.43**	0.63	0.74

Table 13: Feature-wise Macro F1 Scores.

Simulator	Engagement	Verbosity	Self-Disclosure	Resistance
GPT-4.1-mini	–	0.81	0.42	0.15
Llama-3-8B-Instruct	0.06	0.16	0.07	0.15
Qwen-2.5-14B-Instruct	0.09	0.66	0.27	0.10
GPT-5	–	**0.89**	0.42	0.20
DeepSeek-V3.2	0.38	0.85	0.47	0.34
SFT	0.50	0.67	0.51	0.23
Contrastive Learning	0.30	0.55	0.39	0.40
Ours	**0.52**	0.72	**0.58**	**0.42**

Table 14: Feature-wise Pearson correlations. Entries marked as "–" indicate correlations that could not be computed due to zero variance in one of the variables.

Metric	Evaluation Instruction
Linguistic Naturalness	**Definition:** How much the client's utterances sound like a real person in emotional support conversations, rather than language that appears overly refined or AI-generated. **Instruction:** Assess whether the client's language resembles natural, spontaneous human speech in emotional support contexts.
Role Authenticity	**Definition:** How well the speaker stays in the role of a person seeking help, rather than sounding like a counselor, evaluator, or detached observer. **Instruction:** Evaluate whether the client speaks from a first-person, emotionally grounded perspective, without excessive self-analysis or professional terminology.
Psychological Plausibility	**Definition:** Whether the client's emotional reactions and changes over the course of the dialogue feel realistic for a human in a difficult emotional situation. **Instruction:** Examine whether emotional shifts occur gradually and proportionally in response to the conversation, rather than resolving abruptly.

Table 15: Expert evaluation instructions for comparing client utterances across counseling dialogues.

Comparison	**Linguistic Naturalness**			**Role Authenticity**			**Psychological Plausibility**		
	Win	Loss	Tie	Win	Loss	Tie	Win	Loss	Tie
Ours vs. Eeyore	62	19	9	60	20	10	64	13	13
Ours vs. ESC-Judge	62	18	10	65	19	6	56	14	20
Ours vs. ESC-Role	72	9	9	61	16	13	61	12	17

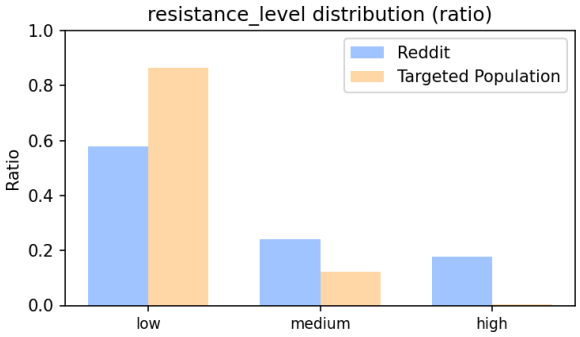
Table 16: Expert evaluation comparing our seeker simulator with existing baselines across three evaluation criteria. Win/Loss/Tie denote the number of instances where our simulator is preferred over, dispreferred to, or tied with the baseline.

Simulator	**Lexical Diversity**			**Semantic Diversity**		**Sentiment Diversity**	
	Distinct-2 ↑	Self-BLEU ↓	TokenRep ↓	APD ↑	ACS ↓	SentMean	SentStd
ClientCAST	0.3031	0.8634	0.5329	0.5832	0.4168	0.9569	0.2619
ESC-Judge	0.3577	0.8389	0.4953	0.5543	0.4457	0.9150	0.3287
ESC-Role	0.4373	0.7882	0.3559	0.7012	0.2988	0.8948	0.3575
Eeyore	0.5201	0.7482	0.2687	0.6767	0.3233	0.6087	0.6454
Ours	**0.5520**	**0.6901**	**0.2186**	**0.7433**	**0.2567**	-0.0555	0.7351

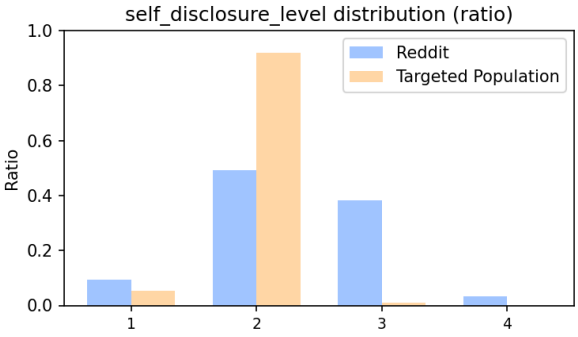
Table 17: Lexical, semantic, and sentiment diversity of seeker utterances across simulators. Arrows indicate whether higher (↑) or lower (↓) values are better. Sentiment statistics are computed using VADER compound scores.

Supporter Model	Seeker Simulator	ES-Skills					General-Skills				
		Iden.	Comf.	Sugg.	Expe.	Info.	Cons.	Role.	Expr.	Huma.	Over.
GPT-5-mini	ClientCAST	4.983	4.990	3.913	2.097	3.563	5.000	5.000	5.000	5.000	5.000
	ESC-Judge	4.980	4.977	4.070	2.473	3.853	5.000	5.000	5.000	5.000	5.000
	ESC-Role	4.987	4.977	3.530	2.033	3.163	4.990	4.990	4.993	4.993	4.990
	Eeyore	4.907	4.940	3.567	2.043	3.503	5.000	5.000	4.997	4.997	4.997
	Ours	**4.410**	**4.477**	**2.820**	**1.367**	**2.393**	**4.933**	**4.963**	**4.860**	**4.830**	**4.853**
Llama-3-8B-Instruct	ClientCAST	4.527	4.967	4.133	2.463	4.030	4.987	4.997	4.990	4.987	4.990
	ESC-Judge	4.750	4.913	4.000	2.333	3.773	5.000	5.000	5.000	5.000	5.000
	ESC-Role	4.421	4.803	4.184	1.960	4.074	5.000	5.000	5.000	5.000	5.000
	Eeyore	4.301	4.816	3.870	1.977	3.836	5.000	5.000	4.993	4.993	4.993
	Ours	**4.097**	**4.297**	**3.300**	**1.590**	**2.970**	**4.883**	**4.920**	**4.860**	**4.813**	**4.840**
Llama-CounselChat	ClientCAST	4.113	4.563	3.960	2.367	4.010	4.947	4.933	4.767	4.737	4.760
	ESC-Judge	4.167	4.447	3.990	2.190	3.830	4.993	4.993	4.897	4.857	4.900
	ESC-Role	4.053	4.227	4.313	2.057	4.313	4.990	4.983	4.850	4.793	4.853
	Eeyore	3.893	3.850	3.820	1.830	4.060	4.803	4.780	4.530	4.313	4.503
	Ours	**3.750**	**3.577**	**3.530**	**1.563**	**3.600**	**4.560**	**4.597**	**4.257**	**3.960**	**4.220**
Llama-ESConv	ClientCAST	3.900	4.220	3.620	2.383	3.353	4.900	4.880	4.640	4.587	4.653
	ESC-Judge	4.000	4.267	3.203	2.587	2.717	4.947	4.960	4.797	4.757	4.827
	ESC-Role	4.053	4.367	3.910	2.320	3.540	4.967	4.963	4.810	4.787	4.833
	Eeyore	3.817	3.927	3.543	2.053	3.133	4.907	4.890	4.587	4.457	4.593
	Ours	**3.390**	**3.150**	**2.303**	**1.213**	**1.807**	**4.387**	**4.363**	**3.960**	**3.617**	**3.887**
Llama-ExTES	ClientCAST	4.427	4.897	3.993	2.257	3.997	4.993	4.993	4.983	4.973	4.980
	ESC-Judge	4.523	4.870	3.797	2.267	3.753	4.997	5.000	4.997	4.997	4.997
	ESC-Role	4.480	4.773	4.140	2.040	4.077	5.000	5.000	5.000	5.000	5.000
	Eeyore	4.333	4.660	3.720	1.903	3.797	5.000	5.000	4.990	4.980	4.990
	Ours	**4.083**	**4.113**	**3.210**	**1.580**	**3.040**	**4.883**	**4.923**	**4.813**	**4.740**	**4.810**
Llama-Psych8k	ClientCAST	4.237	4.720	3.977	2.143	3.937	4.993	4.993	4.977	4.960	4.973
	ESC-Judge	4.277	4.687	4.030	2.297	3.960	5.000	5.000	4.997	4.997	4.997
	ESC-Role	4.227	4.540	4.253	2.090	4.167	5.000	5.000	5.000	5.000	5.000
	Eeyore	4.097	4.183	3.823	1.850	3.843	4.977	4.983	4.923	4.893	4.927
	Ours	**4.000**	**3.803**	**3.573**	**1.643**	**3.517**	**4.867**	**4.883**	**4.733**	**4.587**	**4.710**
Llama-PsyInsight	ClientCAST	4.070	4.243	3.617	1.743	2.847	4.970	4.967	4.900	4.870	4.900
	ESC-Judge	4.297	4.323	3.053	1.750	2.543	4.970	4.993	4.890	4.877	4.917
	ESC-Role	4.167	4.493	3.897	1.940	3.443	5.000	5.000	4.977	4.963	4.977
	Eeyore	3.837	3.563	3.040	1.433	2.537	4.870	4.790	4.653	4.457	4.623
	Ours	**3.503**	**3.017**	**1.983**	**1.087**	**1.547**	**4.418**	**4.431**	**4.167**	**3.843**	**4.070**

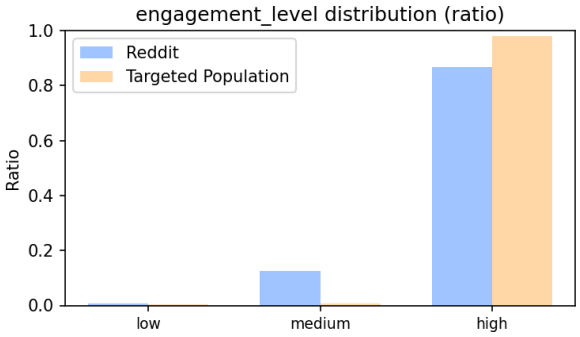
Table 18: Merged results across supporter models and seeker simulators. Scores are averaged and rounded to three decimals. Within each supporter, the lowest score per metric is boldfaced.



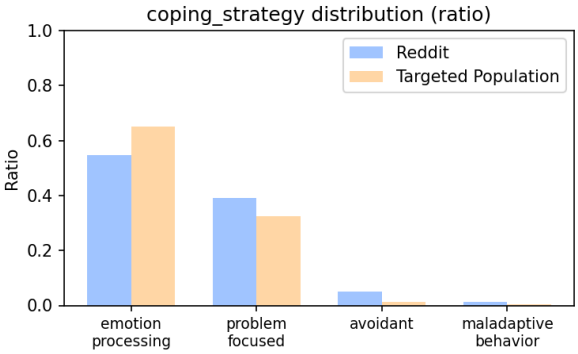
(a) Resistance level distribution



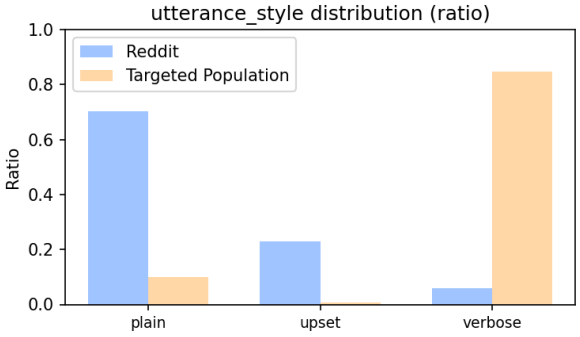
(b) Self-disclosure level distribution



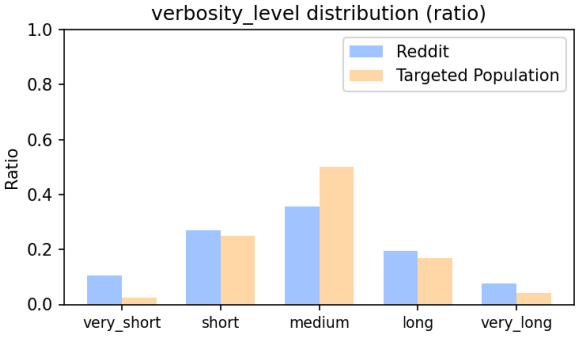
(c) Engagement level distribution



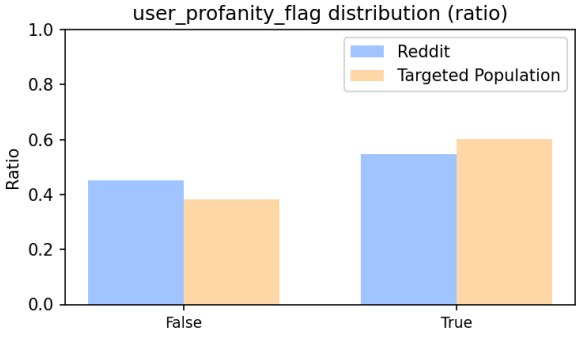
(d) Coping strategy distribution



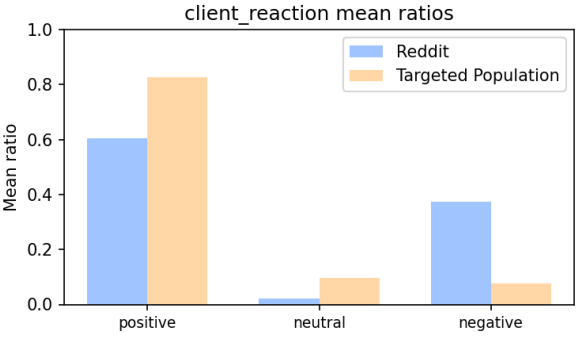
(e) Utterance style distribution



(f) Verbosity level distribution



(g) Profanity flag distribution



(h) Client reaction ratio (means)

Figure 10: Comparison of seeker feature distributions between a target seeker population and our Reddit-based training data.