FIRST, DO NO HARM: AI SUPERVISOR SCAFFOLDS NOVICE GROWTH IN COUNSELOR EDUCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The most dangerous mistakes a novice counselor makes are not the obvious ones: they are utterances that sound caring while quietly violating professional ethics and leaving vulnerable clients less protected. We build an AI supervisor that does not replace novice counselors, but grows them—teaching them to internalize ethical violations they would otherwise never notice. What makes this supervisor non-trivial is not detection but teaching: it must locate the ethical-violating utterance, diagnose the ethical violation against APA principles, and deliver feedback that explains not just what went wrong, but why it is risky and how to respond differently. The core obstacle is that (1) ethical violations are by nature unlabeled in real clinical data, and (2) existing AI counselors trained only to match correct answers will never learn to teach. We resolve both at once: a controllable AI novice that intentionally enacts predefined mistake categories makes supervision labels a natural byproduct of generation, yielding ETHICSCAFF, a 9,915-instance human-in-the-loop dataset; and GRPO under a Novice Growth Reward (NGR) optimizes the supervisor not for answer correctness but for whether a weaker novice model actually improves after reading its explanation. Across patient-disjoint tests, our best models reach 94.37% F1 for violated-principle classification and 74.24% F1 for violation localization; GRPO further improves localization precision, enabling the supervisor to surface context-dependent ethical risks that keyword or retrieval baselines can miss. Expert ratings favor our teaching-oriented explanations for professional depth and actionability, while supervised novices improve across all six counseling metrics. In a four-week pilot, our feedback produces the highest self-efficacy trajectory among all conditions, suggesting guidance that helps learners turn detected problems into safer responses.

1 INTRODUCTION

AI for Novice Counselor Education The field of mental health faces a persistent imbalance between growing demand and the limited availability of trained counselors WHO (2023). Although large language models (LLMs) have been explored as low-cost providers of mental health support Xiao et al. (2024); Na (2024); Xie et al. (2025); Wang et al. (2025c); Shen et al. (2024), deploying them directly in patient-facing treatment remains ethically and clinically fragile. Patient-facing LLMs raise concerns about professional responsibility, safety, privacy, and therapeutic alliance Ong et al. (2024); Raile (2024); Hager et al. (2024); Chung et al. (2023); Nichol et al. (2023); Choudhury & Chaudhry (2024); Scholich et al. (2025). A safer and more scalable role for AI is therefore to accelerate counselor education—the systematic training process that prepares novices to master therapeutic techniques and treatment strategies before working with real clients Wang et al. (2024); Akkurt et al. (2025); Beeson et al. (2025); Wang et al. (2025a); Li et al. (2024); Brügge et al. (2024).

"First, Do No Harm" Prior work expands practice in therapeutic techniques but often underemphasizes ethics, the field's most fundamental obligation. This foundational principle Smith (2005); Sokol (2013) underscores why ethics education must take priority in counselor training. Well-intentioned novices can cause harm they cannot yet recognize Association (2002); Vybíral et al. (2024); Gerke et al. (2020). Consider a novice counselor who, hearing a client from a collectivist culture express guilt about prioritizing family over personal goals, responds: "You need to put yourself first. That's healthy boundaries." The advice sounds empowering, yet it imposes individualistic values, invalidates the client's cultural identity, and may deepen their distress. Similarly, when a

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

client hints at passive suicidal ideation, saying "Sometimes I wonder if people would be better off without me," a novice might respond: "You don't need to think that way. I'm sure your family loves you." This sounds caring, but it is premature reassurance that silences disclosure and bypasses critical risk assessment. The danger is not that novices intend harm, but that they cannot yet recognize it.

AI Supervision Fits as Ethical Learning Scaffold How do novices learn to internalize ethical principles? For novices, learning what not to do is easier than mastering what to do. Expert-level judgment (Fig. 1c) requires weighing countless contextual factors and varies by approach and counselor style, resisting any fixed standard Leichsenring & Steinert (2017); Kramer (2020); Leichsenring et al. (2018). Ethical constraints, by contrast, are more tractable (Fig. 1b): the American Psychological Association (APA) provides a finite set of universal principles, and violations follow recognizable patterns Association (2002); Vybíral et al. (2024); Gerke et al. (2020). Traditional supervision relies on repeated exposure under expert guidance, but is slow, resource-intensive, and inconsistent. The core challenge is what Vygotsky termed the Zone of Proximal Development (ZPD) Vygotsky (1978); Wood et al. (1976): novices cannot see their violations alone yet can learn to recognize them with guidance (Fig. 1a → Fig. 1b). Therefore, we propose an ethical learning framework where simulated patients enable repeated exposure to novices' own mistakes, while an AI supervisor surfaces hidden violations and delivers feedback on where they occur, which principles are breached, and why.

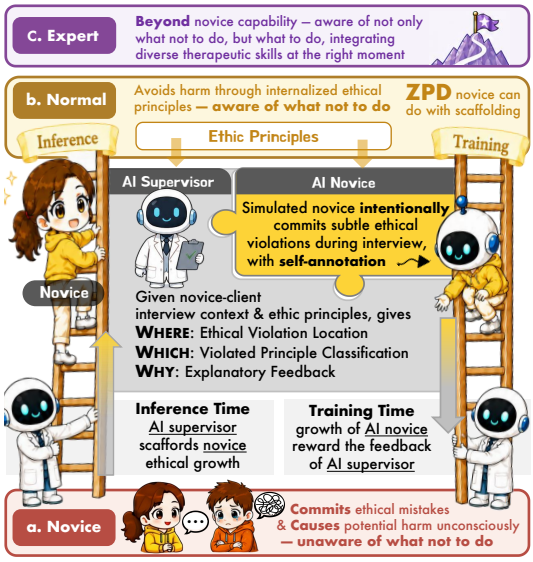


Figure 1: Illustration of how our AI supervisor supports ethical learning within the zone of proximal development (ZPD) and helps novices grow into normal ethical practice. A novice counselor may initially make ethical mistakes without recognizing them, but with guidance can gradually improve. During training, a simulated novice intentionally makes mistakes and learns from the supervisor's "scaffold" feedback; during inference, the trained supervisor reviews a counseling dialogue and provides structured feedback.

Two Obstacles for Training AI Supervisors *First, there is no suitable training data.* Existing open-source counseling datasets Lee et al. (2024); Liu et al. (2023) were curated to train high-quality counselors, not to teach novices what not to do. Their dialogues are expert-validated, ethically vetted, and filtered to exclude the mistakes we need novices to recognize. Safety-focused datasets Ji et al. (2024); Ganguli et al. (2022), while containing harmful content, target overt violations (e.g., abuse, discrimination) rather than the nuanced ethical missteps that counseling novices typically make. More critically, even if subtle violations were present, these datasets contain no ground-truth annotations for supervision. There are no labels indicating where violations occur or which ethical principles are compromised. Most importantly, there are no explanations of why the identified where and which constitute violations, or how to respond differently. Such explanatory scaffolding is essential for human learners to internalize ethical principles, not merely for models to eliminate risks and respond safely to replace human counselors Chung et al. (2023); Scholich et al. (2025); Raile (2024). The supervision task demands both dialogues containing subtle ethics violations and rich pedagogical annotations. Yet existing datasets provide neither.

Second, there is a mismatch of training objectives. Even if suitable data existed, the standard training objective is fundamentally misaligned with the goal of supervision. Existing datasets optimize AI counselors to generate correct responses, evaluated by matching against gold-standard answers Papineni et al. (2002); Lee et al. (2024); Liu et al. (2023)—these works well for training AI counselors, but not AI supervisors. A supervisor's role is not to provide correct answers, but to help novices internalize principles so they can recognize and avoid violations in future cases, rather than merely memorizing corrections for cases they have already encountered. For instance, directly telling a novice "line 3 violates the non-judgment principle" is perfectly "correct" feedback, yet it fails to

explain why this principle matters or how violations harm clients. A pedagogically effective supervisor, returning to the earlier case, would instead explain: "When you said 'You don't need to think that way. I'm sure your family loves you,' this premature reassurance violates non-judgment because it dismisses the client's lived experience and closes the space for risk assessment. The client may interpret your response as 'my feelings aren't valid here' and stop sharing suicidal thoughts." This is the distinction between teaching to the test and teaching to think Jensen et al. (2014); Zakharov & Carnoy (2021).

Our Approach and Contributions Motivated by these challenges, we make three contributions. **(1)** We formulate an ethics-centered supervision task for novice counselor education and construct ETHICSCAFF, a 9,915-instance human-in-the-loop dialogue-feedback dataset. It provides structured supervision targets for locating problematic utterances, classifying novice-counselor mistakes, and explaining why the responses are ethically risky and how they can be improved, using an expert-operationalized taxonomy grounded in professional ethics. By requiring explicit localization and classification before explanation, this formulation can expose subtle, context-dependent ethical risks that surface-level assessment may overlook. **(2)** We introduce a two-stage SFT–GRPO training framework with a Novice Growth Reward (NGR). NGR rewards explanations that improve a weaker novice's localization and classification performance after reading the feedback, evaluating explanations through measurable novice improvement rather than answer matching alone. Together, SFT, GRPO, and NGR yield teaching-oriented explanations that make risks understandable and guide novice counselors toward safer responses, rather than merely stating the answer. **(3)** We evaluate the framework through patient-disjoint simulated test cases, downstream counselor-behavior assessment, automatic and expert feedback evaluation, component ablations, and a four-week study of self-reported self-efficacy among 36 counseling-psychology master's students, showing how learner-oriented optimization can turn detection into guidance that makes context-dependent risks understandable and actionable.

2 RELATED WORKS

Simulated Clients for Mental-Health Training Recent work has explored simulated clients and expert-guided role-play for mental-health education. PATIENT-Ψ Wang et al. (2024) uses LLMs to simulate patients for training mental-health professionals, while Roleplay-doh Louie et al. (2024) enables domain experts to specify principles that guide simulated-patient behavior. These studies primarily focus on constructing realistic and controllable client-side interactions. Our framework uses simulated clients as the training environment, but focuses on the complementary supervision direction: a controllable novice counselor is prompted to exhibit predefined ethical mistakes, and an AI supervisor identifies the problematic utterances, classifies the underlying violations, and provides structured explanatory feedback.

LLM Critique and Supervisory Feedback Recent work has studied LLMs as critics rather than only generators, including benchmarks for critique quality Lin et al. (2024); Lan et al. (2024b;a); Xu et al. (2026); Zhang et al. (2025), evaluator-specialized models such as Prometheus Kim et al. (2023; 2024), and reinforcement-learning methods for improving critique behavior Xi et al. (2025). Feedback generation has also been investigated in counselor-training settings Chaszczewicz et al. (2024). These methods mainly assess or improve the quality of generated critiques, whereas our supervisor must additionally localize ethical violations, classify novice mistakes, and explain how the counselor can improve within a counseling dialogue.

Teacher–Student Learning from Feedback Another line of work evaluates feedback by its effect on a learner. Stronger or privileged teachers have been used to provide supervision for weaker models Choudhury & Sodhi (2025), simulated students have been used to train instructors Markel et al. (2023), and tutoring systems have been optimized for learner improvement across interactions Scarlatos et al. (2025); Demszky et al. (2024). Our work differs in using a Novice Growth Reward: the supervisor is optimized according to whether a weaker novice improves its ethical-violation localization and classification after reading the feedback, rather than the critique quality alone.

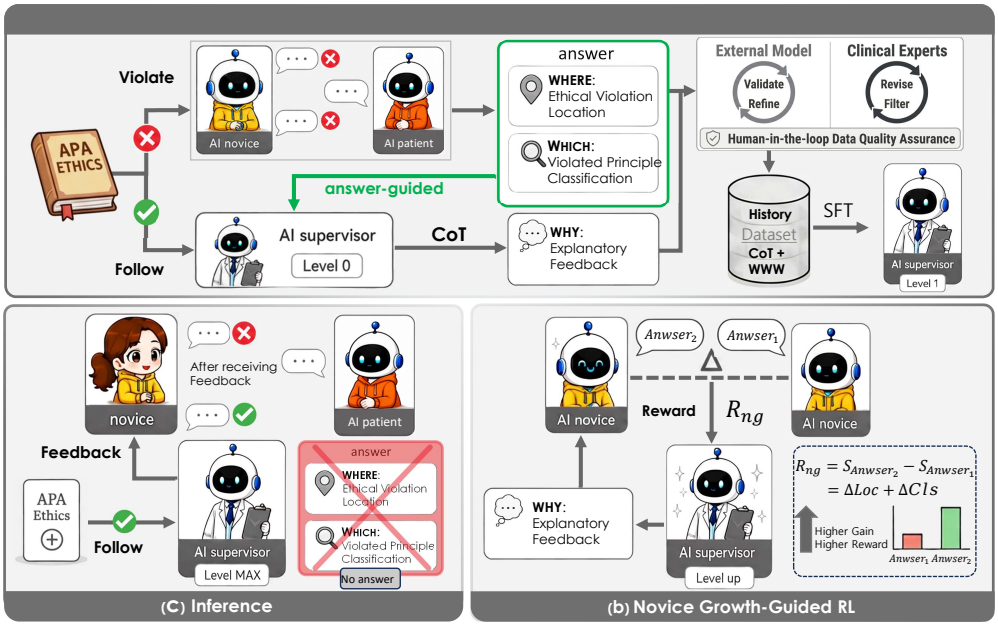


Figure 2: Overview of the proposed framework for ethical-mistake data synthesis and novice supervision. (a) Ethics-violating novice data synthesis builds ETHICSCAFF from controllable novice–patient dialogues, where a novice model intentionally enacts predefined, subtle ethical mistakes, making localization and category labels natural byproducts of generation. A Level-0 supervisor provides answer-guided WHERE–WHICH–WHY supervision, which is refined by external validation and clinical-expert review for SFT initialization of a Level-1 supervisor. (b) Novice Growth-Guided RL further trains the supervisor with GRPO, rewarding explanations that improve a weaker novice's localization and classification without merely disclosing the answer. (c) At inference time, without answer-guided supervision, the trained supervisor has learned to bridge abstract ethical principles and concrete counseling practice by contextualizing them within the dialogue, turning hard-to-notice mistakes into scaffolded guidance for ethical reflection and self-correction.

3 METHODOLOGY

Task Definition Psychological counseling supervision is not merely error detection, but a mechanism for ethical growth. We therefore formulate an LLM-based psychological counseling supervision mechanism that observes counselor-client dialogues and provides professional feedback for novice counselors. Let $\mathcal{H} = \{u_1, u_2, ..., u_n\}$ denote the dialogue history, where u_i represents the utterance in the i-th turn. Let $\mathcal{S}$ denote the set of mistaken utterances, $\mathcal{M}$ the set of mistake categories, and $\mathcal{F}$ the set of feedback. We formalize the supervision task as the following joint probability distribution $P(f, m, s|\mathcal{H})$, where $f \in \mathcal{F}$, $m \in \mathcal{M}$, and $s \in \mathcal{S}$. According to the chain rule of probability, this joint distribution can be decomposed into three sequential conditional probabilities. First, the model performs Ethical Violation Location by computing $P(s|\mathcal{H})$, which locates the specific utterance containing the counselor's mistake based on the dialogue history. Second, given the identified mistaken utterance, the model conducts Violated Principle Classification through $P(m|\mathcal{H}, s)$, determining the clinical category of the identified problem. Finally, the model generates explanatory pedagogical feedback via Feedback Synthesis, modeled as $P(f|\mathcal{H}, s, m)$, which synthesizes all available information to produce actionable guidance. Therefore, the complete generation process can be expressed as:

$$P(f, m, s|\mathcal{H}) = P(s|\mathcal{H}) \cdot P(m|\mathcal{H}, s) \cdot P(f|\mathcal{H}, s, m) \tag{1}$$

This decomposition mirrors the three core steps of real clinical supervision: locating the problematic utterance, identifying the underlying violation, and delivering constructive feedback.

3.1 ETHICS-VIOLATING NOVICE DRIVEN DATA SYNTHESIS

Given the scarcity of high-quality supervision data in psychological counseling, together with privacy constraints, sample imbalance in real clinical data, and the fact that ethical violations are often subtle and rarely labeled, we need a controllable way to make such failures observable and teachable. We therefore collaborate with clinical experts, who distilled 15 common novice-counselor violation types from the *APA Principles* Association (2002), and add one "Exemplary Practice" category to construct Counselor Ethical Principles $\mathbb{P} = \{P_1, P_2, ..., P_{16}\}$, where each principle P_k codifies a novice-relevant failure mode together with its corrective supervision logic as a triple: $P_k = (cat_k, beh_k, strat_k)$ where cat_k denotes the category name, beh_k describes the violation behavior, and $strat_k$ specifies the corresponding supervision explanatory strategy; together, these principles define both what should be corrected and what ethically grounded practice looks like. Table 1 shows one representative example. We then use these principles to build an ethical-violation-driven data synthesis framework that generates controlled dialogue-feedback pairs in which hidden ethical failures become visible, teachable, and learnable.

3.1.1 ROLE-CONDITIONED BEHAVIOR SIMULATION

Table 1: Example of Counselor Ethical Principles

Item	Content
Category	Cultural Insensitivity
Ethics-violating Behavior	The counselor overlooks how the client's culture, gender, religion, or other identities shape their distress, leading to biased understanding or inappropriate intervention.
Explanatory Strategy	Encourage cultural humility: ask how cultural factors shape the client's worldview and distress, avoid assumptions, and treat respect for diversity as a core professional value.

The framework is driven by two independent sets: the Counselor Ethical Principles $\mathbb{P}$ and the Client Case Set Wang et al. (2024) $\mathbb{C} = \{C_1, C_2, ..., C_{106}\}$. Together, they let us ground supervision in professional ethical constraints while preserving realistic variation in client context. For the ethics-violating novice counselor T_k, we instantiate a specialized counselor simulator powered by carefully designed prompts for each $P_k \in \mathbb{P}$. This simulator is prompted to consistently exhibit the k-th specific violation during conversations, creating the kind of novice error that a supervisor must learn to catch and teach through. Its behavior in the i-th dialogue turn is modeled as: $u_i^T = T_k(\mathcal{H}_{i-1}, cat_k, beh_k)$, where u_i^T is the counselor's utterance in the i-th turn, $\mathcal{H}_{i-1} = \{u_1, ..., u_{i-1}\}$ represents the dialogue history up to turn $i-1$, and cat_k and beh_k extracted from P_k determine the simulator's behavior pattern. Correspondingly, for each client case $C_j \in \mathbb{C}$, we instantiate a behavior-sensitive simulated client C_j. This simulator possesses unique background and personality traits while being capable of producing contextually appropriate responses to the counselor's specific mistakes. This interaction is crucial because many ethical violations only become legible when their downstream effect on the client is placed in context. Its behavior is modeled as: $u_i^C = C_j(\mathcal{H}_{i-1} \oplus u_i^T, Case_j)$,where u_i^C is the client's response, $\oplus$ denotes sequence concatenation, $Case_j$ contains the client's information.

3.1.2 DIALOGUE GENERATION AND FEEDBACK SYNTHESIS

In the dialogue generation stage, the system orchestrates multi-turn interactions between an ethics-violating novice counselor T_k and a simulated client C_j, producing a complete dialogue history:

$$D_{j,k} = G_{dialogue}(T_k, C_j) = \{(u_1^T, u_1^C), ..., (u_n^T, u_n^C)\} \tag{2}$$

Following dialogue generation, a LLM-based clinical supervisor $\mathcal{S}$ with an "omniscient perspective" reviews the entire dialogue process $D_{j,k}$ and generates structured critical feedback based on predefined violation definition P_k:

$$F_{j,k} = \mathcal{S}(D_{j,k}, P_k) = (s_{j,k}, m_{j,k}, f_{j,k}) \tag{3}$$

where the feedback $F_{j,k}$ comprises three components: $s_{j,k}$ identifies the problematic sentence by precisely quoting the specific utterance embodying the violation; $m_{j,k}$ denotes the violation category, explicitly stating the category of error committed by the novice counselor; and $f_{j,k}$ provides concrete and actionable feedback such as response strategies and improvement suggestions. In this sense, the synthesized target is not merely a label, but a compact instance of supervision: what went

wrong, why it matters, and how the novice can improve. Finally, the dialogue and feedback are combined into a complete data pair $P_{j,k} = (D_{j,k}, F_{j,k})$, which collectively forms the dataset required for training model. Specifically, we adopt a full cross-combination strategy over 106 client types and 16 counselor-behavior patterns, and sample each combination 6 times with DeepSeek-V3, yielding 10,176 automatically annotated instances with both violation category labels and problematic utterance labels. This synthesis process therefore supplies not only scale, but also the pedagogically structured supervision signal needed to train an AI supervisor for novice growth.

3.2 Data Quality Assurance

From the initial records, we remove 261 samples that fail all four expert-defined quality criteria, retaining 9,915 samples. GPT-4o then refines and validates feedback in a closed loop across progressivity, actionability, ethicality, and supportiveness. Of the retained samples, 9,386 pass without refinement, 418 pass within three sequential VGR iterations, and the remaining 111 are revised by two hospital-based clinical psychologists with doctoral degrees through consensus review. Two clinical psychologists jointly reviewed 400 stratified samples for label consistency and feedback quality. Disagreements were resolved through discussion, and all reviewed samples were approved after consensus correction; further details are provided in Appendix E. Five psychology master's students audited sampled training data for category-name/ID, utterance-ID, verbatim, and semantic leakage; none was found.

3.3 Supervisor Model Training

We train the supervisor in two stages. We first perform SFT on EthicScaff so that the model learns the supervision format and acquires the basic ability to localize ethical-violating utterances, classify mistake types, and provide targeted feedback. We then apply RLVR with GRPO, where the key objective is not only to produce correct supervision outputs, but also to generate explanations that are genuinely *teachable*. This stage completes the bidirectional scaffold: the same system that guides novices at inference time is optimized during training for whether its feedback can actually produce novice growth.

Novice Growth-Guided RL. Our central signal is the Novice Growth Reward(NGR). Given the supervisor's explanation, we let a frozen novice model (Llama3.1-8B) solve the ethical violation localization and classification task before and after reading that explanation. The supervisor is rewarded when its explanation improves the novice's performance without directly revealing the answer. In this way, we optimize for feedback that is instructional and transferable, rather than merely answer-matching. The overall reward is:

$$R_{\text{total}} = w_1 \cdot R_{\text{format}} + w_2 \cdot R_{\text{loc}} + w_3 \cdot R_{\text{classify}} + w_4 \cdot R_{\text{ng}}. \tag{4}$$

Let $(\hat{s}^{\,b}, \hat{m}^{\,b})$ and $(\hat{s}^{\,a}, \hat{m}^{\,a})$ denote the novice predictions before and after reading the explanation. We define the Novice Growth Reward as

$$R_{\text{ng}} = \lambda_1 \cdot \big[\text{Prec}(\hat{s}^{\,a}) - \text{Prec}(\hat{s}^{\,b})\big] + \lambda_2 \cdot \big[\text{Acc}(\hat{m}^{\,a}) - \text{Acc}(\hat{m}^{\,b})\big]. \tag{5}$$

This design encourages the supervisor to provide stepwise, non-leaking guidance that helps a weaker learner improve. More implementation details are provided in Appendix D.

4 Experiments: Detection, Teaching Quality, and Learner Outcomes

We evaluate the proposed supervision framework across five aspects: ethical violation detection, component contributions through ablations, teaching-oriented feedback quality, downstream counselor behavior, and preliminary human-facing evidence through novice self-efficacy. All objective results use the patient-disjoint test set described in Appendix E; test patient cognitive profiles and their generated dialogues are never used for training. Specifically, we partition the 106 patient cognitive models into 94 training and 12 disjoint test patient IDs, yielding 8,923 training and 992 test records. All histories and dialogues associated with a patient remain in the same split. Detailed settings and additional results are provided in Appendix C.

4.1 SUPERVISION TRAINING IMPROVES ETHICAL VIOLATION DETECTION

We examine whether supervision-oriented training sharpens ethical judgment by locating problematic counselor utterances and diagnosing mistake categories. Table 2 reveals a clear asymmetry between the two objective tasks. For untuned open-source models, violated-principle classification remains substantially weaker than proprietary baselines, whereas violation localization is relatively more competitive. This suggests localization relies more on general discourse understanding, while fine-grained discrimination requires task-specific supervision. Tables 2 and 3 report mean $\pm$ standard deviation over five independent runs. After incorporating ETHICSCAFF, open-source models improve dramatically on both tasks and overtake proprietary baselines by a large margin. On Task 1, Qwen3-14B (w/ ETHICSCAFF, Full) achieves 94.37% F1-score, far above GPT-4o with RAG (73.12%) and DeepSeek-V3 (46.85%). On Task 2, Qwen3-8B (w/ ETHICSCAFF, Full) reaches the best overall F1-score of 74.24%, together with the strongest Jaccard (63.03%) and EMR (18.40%), while Qwen3-14B (w/ ETHICSCAFF, Full) remains highly competitive at 73.25% F1-score.

These gains show that supervision-oriented training equips models with ethical judgment largely absent from general-purpose prompting or retrieval. The Full variants further improve over SFT-only variants, indicating that NGR strengthens diagnostic competence as well as feedback optimization. Although RAG variants reach higher recall, they over-predict problematic sentences and do not achieve the best F1-score or EMR. Our models provide a better precision–recall trade-off, suggesting that they filter contextually consequential violations rather than flagging atypical wording indiscriminately. This distinction matters for training: over-flagging gives a novice too many weak warnings, whereas selective localization directs attention to the utterance whose context changes its ethical meaning. The gains therefore support a supervisor that makes hidden risks legible without turning every unusual phrase into a violation.

Table 2: Performance comparison on objective evaluation tasks. For each model family (Llama3.1-8B, Qwen3-8B, and Qwen3-14B), the best result within the group is highlighted in **bold**.

Model Family	Variant	Task 1: Violated Principle Classification				Task 2: Ethical Violation Location				
		Accuracy	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Jaccard	EMR
Closed-Source Models										
GPT-4o	Base	51.18 $\pm$0.20	62.69 $\pm$0.37	51.18 $\pm$0.15	45.66 $\pm$0.17	62.19 $\pm$0.46	89.93 $\pm$0.22	69.48 $\pm$0.33	57.63 $\pm$0.38	13.68 $\pm$0.30
	RAG	73.61 $\pm$0.37	81.52 $\pm$0.25	73.61 $\pm$0.13	73.12 $\pm$0.13	61.90 $\pm$0.37	95.17 $\pm$0.46	71.76 $\pm$0.31	59.59 $\pm$0.25	14.10 $\pm$0.28
Claude-Sonnet-4	Base	54.59 $\pm$0.17	65.82 $\pm$0.18	54.59 $\pm$0.28	47.44 $\pm$0.18	59.57 $\pm$0.17	94.67 $\pm$0.40	70.04 $\pm$0.18	58.04 $\pm$0.01	13.14 $\pm$0.25
Open-Source Models										
DeepSeek-V3-685B	Base	52.14 $\pm$0.05	64.04 $\pm$0.25	52.14 $\pm$0.18	46.85 $\pm$0.15	60.71 $\pm$0.13	97.05 $\pm$0.18	71.85 $\pm$0.11	59.66 $\pm$0.10	13.57 $\pm$0.23
Llama3.1-8B	Base	29.39 $\pm$0.07	43.73 $\pm$0.27	29.39 $\pm$0.43	25.96 $\pm$0.32	59.40 $\pm$0.12	94.31 $\pm$0.20	69.91 $\pm$0.11	57.28 $\pm$0.06	10.01 $\pm$0.13
	Few-shot	18.12 $\pm$0.23	22.00 $\pm$0.46	18.12 $\pm$0.31	16.07 $\pm$0.26	59.75 $\pm$0.13	94.92 $\pm$0.36	70.49 $\pm$0.15	57.86 $\pm$0.04	10.57 $\pm$0.25
	RAG	57.22 $\pm$0.20	62.27 $\pm$0.24	57.22 $\pm$0.33	54.98 $\pm$0.21	59.95 $\pm$0.10	**96.49** $\pm$0.22	71.25 $\pm$0.10	58.81 $\pm$0.00	12.30 $\pm$0.26
	Ours(SFT)	**67.96** $\pm$0.39	**63.22** $\pm$0.46	**67.96** $\pm$0.07	**62.22** $\pm$0.25	**63.64** $\pm$0.47	93.55 $\pm$0.46	**72.70** $\pm$0.37	**61.19** $\pm$0.16	**17.13** $\pm$0.33
Qwen3-8B	Base	23.33 $\pm$0.01	44.16 $\pm$0.31	23.33 $\pm$0.45	22.18 $\pm$0.39	54.78 $\pm$0.07	60.76 $\pm$0.19	53.56 $\pm$0.09	40.52 $\pm$0.06	3.87 $\pm$0.39
	Few-shot	33.55 $\pm$0.16	46.13 $\pm$0.01	33.55 $\pm$0.42	31.58 $\pm$0.28	56.25 $\pm$0.48	**95.64** $\pm$0.10	68.00 $\pm$0.38	54.98 $\pm$0.25	9.94 $\pm$0.41
	CoT	50.22 $\pm$0.26	57.03 $\pm$0.16	50.22 $\pm$0.18	44.80 $\pm$0.12	60.98 $\pm$0.35	91.04 $\pm$0.22	69.52 $\pm$0.26	57.38 $\pm$0.40	12.77 $\pm$0.01
	Few-shot + CoT	57.62 $\pm$0.39	58.95 $\pm$0.15	57.62 $\pm$0.43	51.55 $\pm$0.23	63.18 $\pm$0.18	94.06 $\pm$0.18	72.70 $\pm$0.14	61.16 $\pm$0.21	16.70 $\pm$0.34
	RAG	55.02 $\pm$0.33	63.51 $\pm$0.20	55.02 $\pm$0.44	52.02 $\pm$0.27	58.01 $\pm$0.20	94.85 $\pm$0.28	68.84 $\pm$0.17	56.23 $\pm$0.17	11.86 $\pm$0.01
	Ours(SFT)	73.61 $\pm$0.04	82.17 $\pm$0.48	73.61 $\pm$0.35	75.20 $\pm$0.29	55.99 $\pm$0.02	92.26 $\pm$0.16	66.85 $\pm$0.05	55.26 $\pm$0.44	12.29 $\pm$0.18
	Ours (Full)	**88.98** $\pm$0.01	**91.11** $\pm$0.13	**88.98** $\pm$0.10	**88.63** $\pm$0.08	**66.70** $\pm$0.08	91.45 $\pm$0.28	**74.24** $\pm$0.11	**63.03** $\pm$0.14	**18.40** $\pm$0.12
Qwen3-14B	Base	42.31 $\pm$0.09	66.35 $\pm$0.50	42.31 $\pm$0.49	38.67 $\pm$0.40	60.66 $\pm$0.22	93.62 $\pm$0.14	70.26 $\pm$0.17	57.86 $\pm$0.22	12.50 $\pm$0.16
	Few-shot	52.24 $\pm$0.46	61.59 $\pm$0.31	52.24 $\pm$0.46	47.04 $\pm$0.30	60.54 $\pm$0.12	96.41 $\pm$0.40	71.34 $\pm$0.15	59.01 $\pm$0.25	13.46 $\pm$0.44
	CoT	50.94 $\pm$0.47	59.34 $\pm$0.00	50.94 $\pm$0.40	46.92 $\pm$0.23	61.96 $\pm$0.20	91.66 $\pm$0.13	69.88 $\pm$0.15	57.57 $\pm$0.13	12.37 $\pm$0.50
	Few-shot + CoT	64.15 $\pm$0.09	70.99 $\pm$0.13	64.15 $\pm$0.37	59.87 $\pm$0.21	62.48 $\pm$0.49	94.40 $\pm$0.10	71.79 $\pm$0.36	59.87 $\pm$0.43	15.88 $\pm$0.29
	RAG	68.48 $\pm$0.20	77.69 $\pm$0.05	68.48 $\pm$0.46	68.40 $\pm$0.26	61.23 $\pm$0.43	**97.15** $\pm$0.03	72.18 $\pm$0.32	59.89 $\pm$0.02	13.78 $\pm$0.28
	Ours(SFT)	74.33 $\pm$0.34	75.27 $\pm$0.48	74.33 $\pm$0.49	70.39 $\pm$0.34	63.94 $\pm$0.04	93.37 $\pm$0.17	72.99 $\pm$0.06	61.31 $\pm$0.32	17.11 $\pm$0.13
	Ours(Full)	**94.55** $\pm$0.20	**95.21** $\pm$0.44	**94.55** $\pm$0.18	**94.37** $\pm$0.24	**64.62** $\pm$0.40	93.32 $\pm$0.04	**73.25** $\pm$0.28	**61.65** $\pm$0.17	**17.41** $\pm$0.13

4.2 ABLATIONS REVEAL COMPLEMENTARY GAINS IN TEACHING ORIENTED SUPERVISION

Table 3 shows complementary contributions across the training stack. The Full model reaches 88.63% classification F1 and 74.24% localization F1, compared with 44.80% and 69.52% for the baseline. Removing SFT reduces classification F1 to 47.28%, while removing VGR lowers classification and localization F1 to 59.62% and 46.84%, respectively, showing that task grounding and validated supervision data are both essential. CoT also supports fine grained diagnosis, with classification F1 falling to 68.92% without it. Finally, adding NGR raises localization precision from 64.02% to 66.70% and exact match recovery from 15.92% to 18.40% while recall remains above 91%, indicating that learner oriented optimization makes the supervisor more selective and directs novices to contextually consequential violations rather than simply increasing warnings.

Table 3: Ablation of SFT, CoT-augmented SFT, GRPO, Novice Growth Reward (NGR), and Validator-Guided Refinement (VGR), conducted on the Qwen3-8B model. "–" indicates NGR is not applicable without GRPO.

Setting	VGR	SFT	CoT	GRPO	NGR	Task 1: Violated Principle Classification				Task 2: Ethical Violation Location				
						Accuracy	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Jaccard	EMR
Baseline	✗	✗	✗	✗	–	50.22 ±0.21	57.03 ±0.46	50.22 ±0.34	44.80 ±0.28	60.98 ±0.13	91.04 ±0.47	69.52 ±0.18	57.38 ±0.30	12.77 ±0.31
Full	✓	✓	✓	✓	✓	**88.98** ±0.32	**91.11** ±0.02	**88.98** ±0.33	**88.63** ±0.17	**66.70** ±0.05	91.45 ±0.40	**74.24** ±0.15	**63.03** ±0.47	**18.40** ±0.08
w/o NGR	✓	✓	✓	✓	✗	87.71 ±0.35	89.97 ±0.47	87.71 ±0.37	87.21 ±0.30	64.02 ±0.35	94.09 ±0.33	73.17 ±0.27	61.34 ±0.30	15.92 ±0.31
w/o SFT	✓	✗	✗	✓	✓	52.42 ±0.38	59.74 ±0.50	52.42 ±0.16	47.28 ±0.24	60.96 ±0.09	90.23 ±0.33	69.19 ±0.12	57.29 ±0.36	12.78 ±0.09
w/o CoT	✓	✓	✗	✓	✓	70.25 ±0.24	79.77 ±0.24	70.25 ±0.35	68.92 ±0.22	59.51 ±0.39	**96.83** ±0.49	71.17 ±0.33	58.75 ±0.14	12.24 ±0.10
w/o GRPO	✓	✓	✓	✗	–	73.61 ±0.16	82.17 ±0.30	73.61 ±0.37	75.20 ±0.25	55.99 ±0.37	92.26 ±0.35	66.85 ±0.30	55.26 ±0.26	12.29 ±0.05
w/o VGR	✗	✓	✓	✓	✓	54.65 ±0.10	79.00 ±0.20	54.65 ±0.23	59.62 ±0.17	41.71 ±0.34	61.31 ±0.35	46.84 ±0.27	39.29 ±0.48	11.12 ±0.22

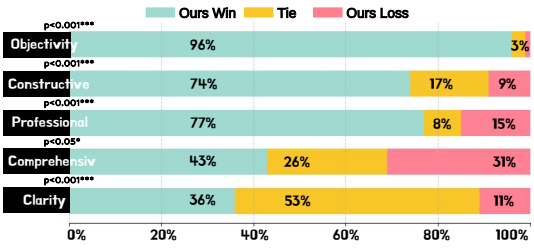


Figure 3: Reward-model evaluation results showing the win, loss, and tie rates of feedback from the ETHICSCAFF-tuned Qwen3-8B relative to feedback from the base model. Results are statistically significant ($p < 0.05$, Bootstrap).

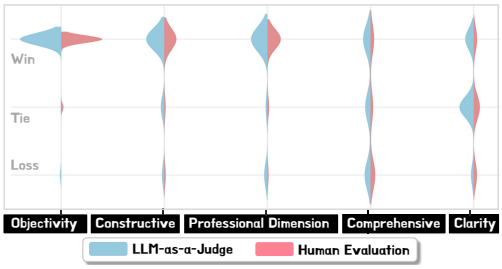


Figure 4: High agreement between LLM and human evaluation in supervisory feedback assessment across evaluated criteria.

4.3 FEEDBACK QUALITY AND TEACHING-ORIENTED SUPERVISION

We examine whether the generated feedback is judged as better supervision. We use Skywork-Reward-Llama-3.1-8B Liu et al. (2024) to assign scalar rewards to feedback from the tuned and base models and derive pairwise preferences on 300 test samples across five criteria: Objectivity & Fairness, Constructiveness & Actionability, Professional Depth, Comprehensiveness, and Clarity & Structure. As shown in Figure 3, fine-tuning with ETHICSCAFF improves critique quality across these dimensions, especially objectivity, professional depth, and constructiveness. Two psychology experts re-conducted pairwise scoring on the same samples, and Figure 4 shows high consistency with the automatic evaluation. The gains in Professional Depth and Constructiveness & Actionability indicate that the feedback is valued for explaining the underlying risk and specifying a corrective next step. These dimensions jointly capture teaching function: fair feedback identifies issue, professional depth supplies rationale, and actionability turns rationale into a revision trainee can practice.

4.4 SUPERVISION IMPROVES DOWNSTREAM COUNSELOR BEHAVIOR

We then examine whether stronger supervision improves downstream counselor behavior. Following the Counselor-Level Evaluation protocol in PsycEval Pan et al. (2026), we compare four counselor conditions, including the novice counselor, the normal counselor instantiated by PsyDT Xie et al. (2025), baseline feedback, and our feedback. Table 4 shows that supervision improves all six psychological metrics, with our ETHICSCAFF-tuned supervisor outperforming the vanilla Qwen-3-8B baseline on every metric. The strongest gains on MITI, PSC, and WAI link feedback quality to perspective-taking, collaboration, and therapeutic alliance rather than surface fluency. We report 90% CIs from 10,000 bootstrap resamples; full setup and metric definitions are provided in Appendix C.2. The improvement pattern is also informative. Relative to the novice counselor, our feedback gains +1.60 on MITI, +1.03 on PSC, and +1.38 on WAI, compared with +0.54, +0.33, and +0.37 under baseline feedback. These gaps indicate that the tuned supervisor changes interactional behavior most strongly where ethical reasoning must be expressed through collaboration and alliance. Its scores also approach the normal counselor on several measures, suggesting that targeted supervision can narrow the novice–expert gap without supplying a fixed response template.

Table 4: Performance Comparison Across Counseling Quality Metrics Under Different Counselor Conditions and Feedback Settings. All metrics are higher-is-better.

Condition	Method	EFT-TFS	HTAIS	MITI	PSC	TES	WAI
No Feedback	Novice Counselor	2.21	4.79	5.62	5.98	7.65	4.68
	Normal Counselor (PsyDT)	**3.01**	5.34	**7.35**	**7.14**	**8.73**	6.00
With Feedback	Baseline Feedback	2.78	5.21	6.16	6.31	7.80	5.05
	Our Feedback	2.97	**5.35**	7.22	7.01	8.54	**6.06**
Improvement over Novice	Baseline Feedback	+0.57	+0.42	+0.54	+0.33	+0.15	+0.37
	Our Feedback	**+0.76**	**+0.56**	**+1.60**	**+1.03**	**+0.89**	**+1.38**

4.5 Preliminary Human-Facing Evidence

Novice self-efficacy pilot. We further test whether automated supervision improves the trainee's own experience by strengthening novice self-efficacy, a key indicator of perceived counseling competence Larson & Daniels (1998); Wang et al. (2025b); Tang et al. (2025). We built an online platform where novice counselling-psychology students conduct weekly text-based sessions with an LLM-driven virtual client and receive feedback over four weeks. Participants were assigned to one of three conditions: no feedback, feedback from a base model, or feedback from our fine-tuned supervisor. Figure 5 visualizes the resulting four-week self-efficacy trajectories.

Thirty-six counselling-psychology master's students (age $M_{\text{age}} = 24$) participated in the study, twelve in each condition. Each participant completed the CASES-R Hahn et al. (2021); Hunsmann et al. (2024) before and after four weekly training sessions. We focused on eight exploration–insight and action items rated from 1 to 6, yielding 1,152 pre–post item pairs. Scores increased in all conditions, with the largest cumulative improvement under our supervisor; the no-feedback and base-supervision groups plateaued later, whereas our method maintained an upward trajectory. This provides preliminary evidence of short-term self-reported learning, not durable clinical skill transfer. The ordering of the trajectories suggests that improvement is related to supervisory quality rather than feedback exposure alone: the base model improves confidence initially, but the tuned supervisor sustains growth across later sessions. This pattern complements the downstream behavior results while remaining exploratory because the pilot has twelve participants per condition and no delayed behavioral measure.

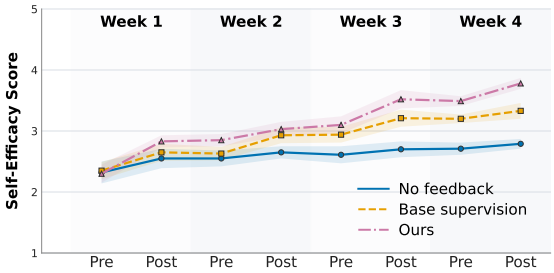


Figure 5: Four-week CASES-R self-efficacy trajectories; shading denotes standard deviation.

5 Discussion & Conclusion

This work reframes mental-health AI as an educational supervisor for scalable novice counselor education. ETHICSCAFF and SFT–GRPO achieve 94.37% classification F1 and 74.24% localization F1; supervised novices improve on all six counseling metrics, and expert ratings favor feedback with greater professional depth and actionability. The results support a complete chain: the model detects a context-dependent violation, explains its consequences, and helps a learner choose a safer response. GRPO's Novice Growth Reward makes this explanation learner-centered, favoring feedback that improves a weaker novice instead of merely naming an error. The patient-disjoint split and ablations provide controlled evidence for these effects, while the pilot offers preliminary evidence of short-term self-reported learning. The downstream results clarify the educational contribution. A fluent response may still ignore cultural context, weaken informed choice, or cross a professional boundary; our feedback makes such hidden risks visible and links them to a principle and a practicable revision. Stronger actionability is followed by gains in perspective-taking, collaboration, therapeutic alliance, and self-efficacy, consistent with a scaffold in which trainees notice a problem, understand why it matters, and practice a safer response. The results suggest a path from detection to transferable ethical judgment, while validation still requires real clinical data, diverse learners, licensed supervisors, and delayed behavioral measures.

REPRODUCIBILITY STATEMENT

We describe the proposed supervision pipeline, including the data-construction framework, the Validator-Guided Refinement protocol, the SFT objective, and the GRPO configuration for the Novice Growth Reward, in Section 3 and Appendix D. The training setup and all hyperparameters for supervision training, inference, and agent-based generation are collected in Appendix C (Tables 5 and 6), and the evaluation protocols for downstream counselor behavior, ethical-violation classification and localization, and the ablations are described in Section 4 and Appendix C. The size, composition, and construction details of ETHICSCAFF are reported in Appendix E, the complete prompt templates in Appendix H, and the expert-defined ethical-principle taxonomy underlying all supervision labels in Appendix G. Anonymous code and sample data, together with documentation for reproducing the training and evaluation pipeline, are provided in the supplementary materials; the code is released under the Apache 2.0 License, and, as detailed in our Ethics Statement, the full dataset is available under controlled research access.

REFERENCES

Mehmet Akkurt, Rakesh K. Maurya, and Timothy Brown. Learning through simulation: Counselor trainees' interactions with ChatGPT as a client. *Behavioral Sciences*, 15(12):1660, 2025. doi: 10.3390/bs15121660.

American Psychological Association. Ethical principles of psychologists and code of conduct. *American Psychologist*, 57(12):1060–1073, 2002. doi: 10.1037/0003-066X.57.12.1060.

Judith S Beck. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications, 2020.

Eric Tod Beeson, Yusen Zhai, Russell Fulmer, Andrew M. Burck, and Rakesh Maurya. A pilot study evaluating the fidelity of ChatGPT in client simulations. *Journal of Counselor Preparation and Supervision*, 19(3):1–16, 2025. doi: 10.70013/d2a58d79.

Emilia Brügge, Sarah Ricchizzi, Malin Arenbeck, Marius Niklas Keller, Lina Schur, Walter Stummer, Markus Holling, Max Hao Lu, and Dogus Darici. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: A randomized controlled trial. *BMC Medical Education*, 24(1):1391, 2024. doi: 10.1186/s12909-024-06399-7.

Alicja Chaszczewicz, Raj Shah, Ryan Louie, Bruce Arnow, Robert Kraut, and Diyi Yang. Multi-level feedback generation with large language models for empowering novice peer counselors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4130–4161, 2024. doi: 10.18653/v1/2024.acl-long.227. URL https://aclanthology.org/2024.acl-long.227/.

Avishek Choudhury and Zaira Chaudhry. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. *Journal of Medical Internet Research*, 26:e56764, 2024.

Sanjiban Choudhury and Paloma Sodhi. Better than your teacher: LLM agents that learn from privileged AI feedback. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=st7XqFgbAH.

Neo Christopher Chung, George Dyer, and Lennart Brocki. Challenges of large language models for mental health counseling. *arXiv preprint arXiv:2311.13857*, 2023.

Suzanne E Decker, Charla Nich, Kathleen M Carroll, and Steve Martino. Development of the therapist empathy scale. *Behavioural and cognitive psychotherapy*, 42(3):339–354, 2014.

Dorottya Demszky, Jing Liu, Heather C. Hill, Dan Jurafsky, and Chris Piech. Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, 46(3):483–505, 2024. doi: 10.3102/01623737231169270. URL https://doi.org/10.3102/01623737231169270.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Leonie Gerke, Ann-Katrin Meyrose, Inga Ladwig, Winfried Rief, and Yvonne Nestoriuc. Frequencies and predictors of negative effects in routine inpatient and outpatient psychotherapy: two observational studies. *Frontiers in psychology*, 11:2144, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

Daniela Hahn, Florian Weck, Michael Witthöft, and Franziska Kühne. Assessment of counseling self-efficacy: validation of the german counselor activity self-efficacy scales-revised. *Frontiers in psychology*, 12:780088, 2021.

Joanna Joy Hunsmann, Destina Sevde Ay-Bryson, Scarlett Kobs, Nicole Behrend, Florian Weck, Michel Knigge, and Franziska Kühne. Basic counseling skills in psychology and teaching: validation of a short version of the counselor activity self-efficacy scales. *BMC psychology*, 12(1): 32, 2024.

Jamie L Jensen, Mark A McDaniel, Steven M Woodard, and Tyler A Kummer. Teaching to the test... or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, 26(2):307–329, 2014.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Juntao Dai, Boren Zheng, Tianyi Qiu, Jiayi Zhou, Kaile Wang, Boxuan Li, Sirui Han, Yike Guo, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. *arXiv preprint arXiv:2406.15513*, 2024.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models, 2024. *URL https://arxiv. org/abs/2310.08491*, 2023.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024.

Ueli Kramer. Individualizing psychotherapy research designs. *Journal of Psychotherapy Integration*, 30(3):440, 2020.

Tian Lan, Wenwei Zhang, Chengqi Lyu, Shuaibin Li, Chen Xu, Heyan Huang, Dahua Lin, Xian-Ling Mao, and Kai Chen. Training language models to critique with multi-agent feedback. *arXiv preprint arXiv:2410.15287*, 2024a.

Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian-Ling Mao. Criticeval: Evaluating large-scale language model as critic. *Advances in Neural Information Processing Systems*, 37:66907–66960, 2024b.

Lisa M Larson and Jeffrey A Daniels. Review of the counseling self-efficacy literature. *The Counseling Psychologist*, 26(2):179–218, 1998.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, et al. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. *arXiv preprint arXiv:2407.03103*, 2024.

Falk Leichsenring and Christiane Steinert. Is cognitive behavioral therapy the gold standard for psychotherapy?: The need for plurality in treatment and research. *Jama*, 318(14):1323–1324, 2017.

Falk Leichsenring, Allan Abbass, Mark J Hilsenroth, Patrick Luyten, Thomas Munder, Sven Rabung, and Christiane Steinert. "gold standards," plurality and monocultures: The need for diversity in psychotherapy. *Frontiers in Psychiatry*, 9:159, 2018.

Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:2404.13066*, 2024. doi: 10.48550/arXiv.2404.13066.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. Criticbench: Benchmarking llms for critique-correct reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1552–1587, 2024.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024.

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

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*, 2024.

Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. GPTeach: Interactive TA training with GPT-based students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, pp. 226–236, Copenhagen, Denmark, 2023. Association for Computing Machinery. doi: 10.1145/3573051.3593393. URL https://doi.org/10.1145/3573051.3593393.

Hongbin Na. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. *arXiv preprint arXiv:2403.16008*, 2024.

Ariadne A Nichol, Meghan C Halley, Carole A Federico, Mildred K Cho, and Pamela L Sankar. Not in my ai: Moral engagement and disengagement in health care ai development. In *Pacific symposium on biocomputing. Pacific Symposium on Biocomputing*, volume 28, pp. 496, 2023.

Jasmine Chiat Ling Ong, Shelley Yin-Hsi Chang, Wasswa William, Atul J Butte, Nigam H Shah, Lita Sui Tjien Chew, Nan Liu, Finale Doshi-Velez, Wei Lu, Julian Savulescu, et al. Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, 6(6): e428–e432, 2024.

Qianjun Pan, Junyi Wang, Jie Zhou, Yutao Yang, Junsong Li, Kaiyin Xu, Yougen Zhou, Yihan Li, Jingyuan Zhao, Qin Chen, et al. Psycheval: A multi-session and multi-therapy benchmark for high-realism and comprehensive ai psychological counselor. *arXiv preprint arXiv:2601.01802*, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

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

Paolo Raile. The usefulness of chatgpt for psychotherapists and patients. *Humanities and Social Sciences Communications*, 11(1):1–8, 2024.

Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. Training LLM-based tutors to improve student learning outcomes in dialogues. In *Artificial Intelligence in Education*, volume 15877 of *Lecture Notes in Computer Science*, pp. 251–266. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-98414-3_18. URL https://doi.org/10.1007/978-3-031-98414-3_18.

Till Scholich, Maya Barr, Shannon Wiltsey Stirman, and Shriti Raj. A comparison of responses from human therapists and large language model–based chatbots to assess therapeutic communication: Mixed methods study. *JMIR Mental Health*, 12(1):e69709, 2025.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*, 2020.

Hao Shen, Zihan Li, Minqiang Yang, Minghui Ni, Yongfeng Tao, Zhengyang Yu, Weihao Zheng, Chen Xu, and Bin Hu. Are large language models possible to conduct cognitive behavioral therapy? In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3695–3700. IEEE, 2024.

Cedric M Smith. Origin and uses of primum non nocere—above all, do no harm! *The Journal of Clinical Pharmacology*, 45(4):371–377, 2005.

Daniel K Sokol. "first do no harm" revisited. *Bmj*, 347, 2013.

Bixia Tang, Jiarong Liang, Wenshuang Hu, and Heng Luo. Enhancing programming performance, learning interest, and self-efficacy: The role of large language models in middle school education. *Systems*, 13(7):555, 2025.

Zbyněk Vybíral, Benjamin M Ogles, Tomáš Řiháček, Barbora Urbancová, and Veronika Gocieková. Negative experiences in psychotherapy from clients' perspective: A qualitative meta-analysis. *Psychotherapy Research*, 34(3):279–292, 2024.

Lev S Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard University Press, 1978.

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

Qikai Wang, Yang Gao, and Xiaochen Wang. Exploring engagement, self-efficacy, and anxiety in large language model efl learning: A latent profile analysis of chinese university students. *International Journal of Human–Computer Interaction*, 41(12):7815–7824, 2025b.

Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. Patient-psi: 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*, pp. 12772–12797, 2024.

Shiquan Wang, Ruiyu Fang, Zhongjiang He, Shuangyong Song, and Yongxiang Li. Emotional support with llm-based empathetic dialogue generation. *arXiv preprint arXiv:2507.12820*, 2025c.

WHO. Mental health: strengthening our response. Technical report, World Health Organization, 2023. URL https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response.

David Wood, Jerome S Bruner, and Gail Ross. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100, 1976.

Zhiheng Xi, Jixuan Huang, Xin Guo, Boyang Hong, Dingwen Yang, Xiaoran Fan, Shuo Li, Zehui Chen, Junjie Ye, Siyu Yuan, et al. Critique-rl: Training language models for critiquing through two-stage reinforcement learning. *arXiv preprint arXiv:2510.24320*, 2025.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. Healme: Harnessing cognitive reframing in large language models for psychotherapy. *arXiv preprint arXiv:2403.05574*, 2024.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. Psydt: Using llms to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1081–1115, 2025.

Chen Xu, Tian Lan, Zhenyu Lv, Qunxi Dong, Jieshuo Zhang, Heyan Huang, Minqiang Yang, and Bin Hu. Bridging the gap between data distribution and model: Dynamic data distribution optimization for improving critique capabilities of large language models. *Expert Systems with Applications*, 300:129878, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv e-prints*, pp. arXiv–2412, 2024.

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

Andrey Zakharov and Martin Carnoy. Does teaching to the test improve student learning? *International Journal of Educational Development*, 84:102422, 2021. doi: 10.1016/j.ijedudev.2021.102422.

Alexander Zhang, Marcus Dong, Jiaheng Liu, Wei Zhang, Yejie Wang, Jian Yang, Ge Zhang, Tianyu Liu, Zhongyuan Peng, Yingshui Tan, et al. Codecriticbench: A holistic code critique benchmark for large language models. *arXiv preprint arXiv:2502.16614*, 2025.

A Limitations

A.1 Simulation-Based Data Realism

EthicScaff is constructed through ethical-violation-driven simulation rather than large-scale real supervision records. This design reflects practical and ethical constraints in counseling education: supervision data are sensitive, difficult to anonymize, and rarely shareable, while recent work has also explored LLM role-play for scalable mental-health training interactions Qiu & Lan (2024); Wang et al. (2024); Louie et al. (2024). To assess generalization across client contexts, we use a patient-disjoint split: 94 of the 106 patient cognitive models are assigned to training and 12 to testing, and every dialogue generated from a patient remains in that patient's split. Thus, the test patient profiles, histories, and generated interactions are unseen during training. However, both splits retain the same expert-defined ethics taxonomy and synthetic generation pipeline, so the generated targets operationalize this fixed taxonomy rather than constitute independently collected clinical gold labels. Accordingly, objective results on this split measure generalization to unseen simulated patient contexts within the construction framework; they do not by themselves establish validity on unseen taxonomies, independently annotated clinical supervision records, or real-world deployment. Simulated dialogues inevitably abstract away some real-world ambiguity, long-term therapeutic history, and institutional context. We reduce this gap through expert-defined principles, Validator-Guided Refinement, and expert review; future work can further validate the framework with de-identified real training cases under appropriate ethical approval. The final expert audit uses joint, non-blind consensus review and therefore serves as a curation safeguard rather than an independent estimate of VGR effectiveness or inter-rater reliability.

A.2 Scope of Ethical Supervision

Our taxonomy focuses on 15 common novice-counselor mistakes grounded in professional ethical principles Association (2002). This scope is intentional: psychotherapy does not have a single universal gold standard across modalities and client contexts Leichsenring & Steinert (2017); Kramer

(2020), whereas ethical constraints are comparatively stable, teachable, and suitable for novice training. The current taxonomy therefore provides an ethics-centered foundation rather than an exhaustive supervision curriculum. Future extensions may incorporate multi-session case formulation, modality-specific intervention timing, and jurisdiction-specific standards.

A.3 Human-Facing Evidence

Our evaluations include objective tasks, downstream counselor-behavior assessment, expert judgment, and a novice self-efficacy study. The human study involves thirty-six counseling-psychology master's students, twelve per condition, and compares no feedback, feedback from a base model, and feedback from our trained supervisor over four weekly sessions. We measure self-efficacy before and after each session using eight CASES-R items, a validated and relevant construct for counselor training Larson & Daniels (1998); Hahn et al. (2021); Hunsmann et al. (2024). These results offer encouraging preliminary evidence of longitudinal self-efficacy growth, while the comparison rests on twelve participants per arm and does not include a post-study follow-up or objective behavioral outcome; larger randomized studies with diverse trainees, licensed supervisors, and delayed assessments would further clarify durable learning effects and downstream clinical relevance.

A.4 Reward Proxy and Deployment Boundary

The Novice Growth Reward uses improvement of a simulated novice model as a tractable proxy for educational usefulness. This design aligns with recent work that optimizes feedback or tutoring through learner-centered outcomes Demszky et al. (2024); Scarlatos et al. (2025), while recognizing that human trainees may differ in background, culture, and supervision needs. Accordingly, our system is intended as a supervised educational scaffold, not an autonomous clinical authority or patient-facing therapist. This boundary is especially important given continuing concerns around safety, workflow integration, professional responsibility, and ethical governance for LLMs in clinical and mental-health settings Ong et al. (2024); Hager et al. (2024); Chung et al. (2023). Future work should integrate human learning signals, personalize feedback, and maintain privacy, auditing, and licensed-supervisor oversight.

B Ethics Statement

This research has been assessed as minimal risk and adheres to established ethical principles for human subjects research. All potential risks have been thoroughly identified with appropriate mitigation strategies in place. The primary risk involves potential mild negative emotional responses among participants when receiving AI-generated supervisory feedback, which are transient and do not exceed the intensity typically encountered during standard professional training or supervision.

Prior to participation, all subjects received comprehensive study information and provided informed consent through an electronic form (Figure 6), confirming voluntary participation with the right to withdraw unconditionally at any time. The research team maintains strict confidentiality of all participant information, with data anonymized and stored securely for research purposes only. In addition, all constructed data used in this study were processed to remove or mask privacy-sensitive personal information, with names replaced to prevent the disclosure of identifiable details. All participants are adults with appropriate counseling backgrounds, ensuring suitable professional context for the training intervention. All participants in the user study were compensated for their time and effort with a remuneration exceeding the national average daily wage.

Because EthicScaff intentionally contains simulated examples of novice counselors committing ethical mistakes, we treat the dataset as a dual-use research artifact. Although these examples are constructed for supervision, detection, and corrective-feedback training, unrestricted release could enable misuse, such as training systems to imitate unsafe counseling behaviors or extracting harmful response patterns for patient-facing use. We therefore adopt a responsible-release strategy. Code, documentation, aggregate statistics, and a safety-screened representative subset are provided to support review and reproducibility, while access to the full dataset is controlled through a research-use request and data-use agreement. The agreement restricts use to counselor education, safety evaluation, supervised research, and model auditing; prohibits patient-facing deployment, attempts to

reproduce harmful counseling behavior as an end goal, and re-identification or reconstruction of sensitive personal details; and requires appropriate institutional oversight where applicable.

Informed Consent Form for Participants

Combining AI and psychological supervision theory, this project aims to provide novice psychological counselors with a platform for skill practice and supervisory feedback. This platform is part of an academic study exploring the effectiveness of AI-assisted psychological counselor training. Please read this informed consent form carefully before use.

Research Purpose

To evaluate the effectiveness, usability, and safety of an AI-assisted supervision system based on large language models. You will engage in simulated counseling with an AI "virtual patient" and receive real-time feedback from an "AI supervisor." We aim to explore this human-AI collaboration's potential to improve counselors' reflection and skills, providing evidence for future training tools.

Participation Process

1. Training and Consent: Brief online training about system, procedures, and confidentiality. You will then give online consent.
2. Baseline Questionnaire: Anonymous online survey on demographics, training, and experience.
3. Simulated Counseling: Text dialogue with an AI "virtual patient" simulating a real session.
4. Feedback Evaluation: Receive an AI-generated report analyzing your performance and suggestions; you evaluate its usefulness.
5. Follow-up Interview: Some participants may be invited for a short online interview about their experience.

The process takes about 30 minutes, all voluntary.

Data Collection and Privacy

- Data include your questionnaire, counseling dialogues, and feedback evaluations.
- All data are anonymized and stored securely, with personal info encrypted separately.
- Data used only for research and system improvement; no personal info disclosed.
- You may request access or deletion of your data anytime.

Possible Risks

- Mild negative emotions like frustration or anxiety may occur but should not exceed normal training experiences.
- Rarely, AI may produce inaccurate or inappropriate content despite filters.
- Online data research carries theoretical privacy risks.
- This platform is not for emergency crisis intervention; stop use and seek professional help if distressed.

Potential Benefits

- Safe environment to practice counseling skills without risk to real clients.
- Personalized, immediate feedback to enhance self-awareness.
- Improved professional reflection through AI's external perspective.
- Exposure to innovative AI applications in mental health training.

Note: Feedback is for reference and cannot replace expert supervision or therapy.

Participation Guidelines

Participation is voluntary; you can withdraw anytime without penalty.
Participants must be 18+ with a counseling background (student, intern, or practitioner).

Please confirm by checking below:
☐ I am over 18, have read and understood this consent form, and voluntarily agree to participate. I understand the risks and benefits and know I can withdraw anytime.

Figure 6: Informed consent form used in the study.

C Detailed Experiments Results

C.1 Training Setup

We train the supervisor in two stages, namely supervised fine-tuning (SFT) and GRPO-based RLVR. In the SFT stage, we use two NVIDIA H800 GPUs(one have 80GB memory) and train on samples formatted as therapist-client dialogue + query + CoT + answer. In the GRPO stage, we use four NVIDIA H800 GPUs, with one GPU dedicated to rollout generation and three GPUs used for policy training; the training samples are formatted as therapist-client dialogue + query + answer. For the 8B model, SFT takes approximately 2 hours, while GRPO training takes approximately 20 hours. The detailed training hyperparameters are summarized in Table 5, and the inference and agent-generation settings are summarized in Table 6.

Our experiments were conducted on open-source models with relatively small parameter counts, including Llama3.1-8B Grattafiori et al. (2024), Qwen3-8B/14B Yang et al. (2025). For baseline comparisons, we evaluated on the open-source Qwen2.5-1.5B/7B Yang et al. (2024), DeepSeek-V3-685B Guo et al. (2025) models and closed-source models GPT-4o and Claude-Sonnet-4.

Table 5: Training hyperparameters for SFT and GRPO.

Hyperparameter	SFT	GRPO
GPUs	2×H800	4×H800 (1 rollout, 3 training)
Optimizer	AdamW	AdamW
Learning rate	2e-4	1e-5
Batch size	16	4
Gradient accumulation steps	16	16
Training epochs	2	1.5
LoRA rank r	16	32
LoRA alpha α	32	64
LoRA dropout	0.05	0.05
Max completion length	4096	8192
Sampling temperature	0.7	0.7
Top-p	0.9	0.9

Table 6: Inference and agent-generation settings.

Hyperparameter	Value
Maximum retry attempts N_{retry}	3
Client agent temperature	0.15
Therapist agent temperature	0.10
Supervisor agent temperature	0

Figure 7 shows the reward trajectories during GRPO training of Qwen3-8B. The total reward increases steadily over training, while the location, format, classify, and explain rewards remain stable and gradually improve, indicating that the policy learns to better satisfy the structured feedback requirements defined in our reward design.

C.2 Details of Downstream Training Effect on Counselor Behavior

We conduct a downstream role-play evaluation to test whether stronger supervision can improve the final behavior of a novice counselor during interaction. Four counselor conditions are compared: (*i*) a novice counselor that directly responds to the simulated patient without external guidance, (*ii*) a normal counselor Xie et al. (2025) that serves as a stronger unguided reference, (*iii*) a supervised novice that revises its draft according to feedback from a baseline supervisor (Qwen3-8B), and (*iv*) our method, where the same novice counselor revises its response under feedback generated by the Qwen3-8B supervisor trained on EthicScaff. This design isolates whether supervision quality can translate into better revised counselor utterances, rather than merely better-looking feedback.

Following the Counselor-Level Evaluation protocol in PsycEval Pan et al. (2026), we evaluate the resulting dialogues from the perspective of clinical proficiency, focusing on whether the counselor delivers theoretically grounded and ethically sound interventions while maintaining longitudinal clinical coherence across different therapeutic modalities. Table 7 summarizes the six metrics used in this evaluation. All metrics are higher-is-better.

The results in Table 4 show three main patterns. First, supervision consistently helps a mistake-prone novice counselor: compared with the unsupervised novice, the baseline supervised novice improves on all six metrics. Second, our supervisor further outperforms the baseline supervisor on every metric, indicating that the advantage does not come only from adding a revise-after-feedback step, but from providing more useful supervision. Third, the gains are particularly large on MITI, PSC, and WAI, where our method improves by +1.60, +1.03, and +1.38 over the novice counselor, respectively. These improvements suggest stronger support for collaborative interviewing, psychodynamic appropriateness, and alliance building.

Another notable pattern is that our method approaches the normal counselor on most dimensions while remaining clearly better than the novice baseline. For example, our system reaches 7.22 on



Figure 7: Reward trajectories during GRPO training of Qwen3-8B. From left to right, we show the total reward, location reward, format reward, classify reward, and explain reward.

Table 7: Definitions of the evaluation metrics used in downstream counselor-behavior assessment.

Metric	**Orientation**	**Definition**
EFT-TFS	Emotion-Focused Therapy	Therapeutic Focus Scale. Evaluates whether the counselor helps the client stay with emotionally salient material and facilitates productive affective processing rather than shifting away from core emotional experience.
HTAIS	General helping process	Helping Transaction Audit Inventory Scale. Measures the quality and integrity of the helping process, including whether the counselor's responses sustain a coherent, facilitative interaction sequence.
MITI	Motivational interviewing / postmodern	Motivational Interviewing Treatment Integrity. Assesses collaborative, non-directive, and change-oriented counseling skill, especially the ability to evoke client reflection without becoming overly controlling or prescriptive.
PSC	Psychodynamic	Psychodynamic Supervision Checklist. Captures the quality of psychodynamic interventions, including sensitivity to underlying conflict, defenses, and relational dynamics emerging in the session.
TES	Humanistic	Therapist Empathy Scale. Measures the degree to which the counselor accurately recognizes, reflects, and deepens the client's emotional experience in an empathic and supportive manner.
WAI	General alliance quality	Working Alliance Inventory. Evaluates the strength of the therapeutic alliance, including agreement on goals, coordination on tasks, and the interpersonal bond formed with the client.

MITI compared with 7.35 for the normal counselor, and slightly surpasses the normal counselor on HTAIS (5.35 vs. 5.34) and WAI (6.06 vs. 6.00). Taken together, these findings indicate that supervision learned from ETHICSCAFF transfers effectively to a realistic training loop and can help weaker counselors revise their responses toward higher-quality therapeutic behavior.

Table 8: Model performance on empathy evaluation. Accuracy and F1 are from empathy classification; TES measures multi-turn empathic quality of supervisor responses, as described above. p-value < 0.05.

Model	Emp. Cls.		TES
	Acc.	F1	Avg. ↑
Closed-Source Models			
GPT-4o	42.7	44.7	8.82
Claude-Sonnet-4	23.6	24.8	8.12
Open-Source Models			
DeepSeek-R1-685B	31.8	34.9	8.45
DeepSeek-V3-685B	29.0	32.7	8.36
Mistral-7B-v0.3	27.9	31.1	8.32
Gemma-3-4B	16.4	9.9	7.82
Gemma-3-12B	22.9	21.4	8.08
Qwen2.5-1.7B	14.8	8.9	7.75
Qwen2.5-7B	25.9	26.5	8.15
Qwen-3-8B	24.7	25.0	8.13
Qwen-3-8B (w/ ETHICSCAFF)	**62.3**	**64.1**	**8.92**
Qwen-3-14B	22.9	21.6	8.08
Qwen-3-14B (w/ ETHICSCAFF)	**74.7**	**76.5**	**9.06**

C.3 EMPATHY CLASSIFICATION AND TES-BASED SUPERVISOR EVALUATION

To examine whether supervision tuning improves broader therapeutic competence, we evaluate models with two complementary empathy-oriented views. First, we retain zero-shot empathy classification on the Empathy-Mental-Health dataset Sharma et al. (2020) (911 samples), which tests whether a response can be recognized as empathic. Second, we adapt the Therapist Empathy Scale (TES) Decker et al. (2014) to evaluate supervisor quality in a multi-turn setting. Although TES was originally developed for human counselors, the same construct is central to high-quality supervision: useful feedback should be grounded not only in mistake detection, but also in accurate understanding of the client's emotional experience and the therapeutic meaning of the current turn.

Following the multi-turn counselor-level evaluation setting of PsycEval Pan et al. (2026), each model is given the full dialogue context and asked to generate the next supportive response. The generated response is then rated by DeepSeek-V3.1 as an automatic judge on a 1–7 Likert scale. We score nine TES-inspired dimensions: acceptance of feelings, attunement to the client's inner world, concern, expressiveness, resonance with client feelings, responsiveness, understanding the client's cognitive framework, understanding feelings, and warmth. We then linearly rescale each dimension score from 1–7 to 0–10 via $\hat{s} = (s-1) \times \frac{10}{6}$, and report the final TES score as the average of the normalized dimension scores. Higher scores indicate stronger empathic resonance and a more growth-promoting supervisory style.

These dimensions probe complementary aspects of supervisory quality. Acceptance and warmth reflect whether the feedback maintains psychological safety; attunement, resonance, and understanding of feelings test whether the model captures the client's latent emotional state rather than responding generically; and responsiveness together with cognitive understanding test whether the suggested next turn remains grounded in the client's meaning structure and the local dialogue trajectory. We therefore use TES to assess whether supervision learned from ETHICSCAFF transfers from error identification to empathic therapeutic judgment.

Table 8 shows a clear pattern for the classification view. Fine-tuning on ETHICSCAFF yields large gains for both model scales and pushes them beyond all baselines, with Qwen-3-14B (w/ ETHICSCAFF) performing best overall and Qwen-3-8B (w/ ETHICSCAFF) also showing strong transfer. Relative to their untuned versions, the 8B model improves by +37.6 Accuracy / +39.1 F_1, and the 14B model by +51.8 / +54.9. Notably, even the fine-tuned 8B model outperforms DeepSeek-R1-685B by 30.5 points in Accuracy. The TES score complements sentence-level classification by testing whether the supervisor can sustain empathic, clinically usable guidance in context.

These results suggest that supervision tuning does more than teach the model to point out mistakes. It also appears to instill therapeutic priors about what constitutes an appropriate, attuned response. This is especially meaningful for novice-support settings: a strong supervisor should not only detect errors after they occur, but also demonstrate the empathic standards that novices are still learning to internalize. In this sense, ETHICSCAFF may improve supervision quality along two complementary dimensions, explicit error correction and implicit modeling of empathic therapeutic judgment.

C.4 MORE OBJECTIVE EVALUATION RESULTS

Reward-model evaluation. We use Skywork-Reward-Llama-3.1-8B Liu et al. (2024) to evaluate the quality of supervisory feedback along the five dimensions defined in Table 9. Because the reward model assigns a scalar reward to a prompt–response conversation rather than directly generating a pairwise judgment, we instantiate one rubric-specific prompt for each dimension. Each prompt contains the therapist–client dialogue, the expert reference answer, and the corresponding rubric from Table 9. The feedback generated by a model is placed in the assistant-response position and scored against this common prompt. Thus, each pair of feedback responses receives five separate reward-score pairs, one for each evaluation dimension.

For every sample and dimension, we form two conversations with identical user prompts: one containing feedback from the ETHICSCAFF-tuned Qwen3-8B and the other containing feedback from the base Qwen3-8B. Before the two conversations are passed to the reward model, their positions in the input list are randomly swapped with equal probability. We retain the model-identity mapping and restore it after scoring, ensuring that the reported comparison is associated with the generating model rather than the first or second input position. Let r_{ours} and r_{base} denote the resulting scalar rewards. We record a tie when $|r_{\text{ours}} - r_{\text{base}}| \leq 3$, because differences within three reward points are treated as insufficiently discriminative. Otherwise, the higher-scoring response is preferred: $r_{\text{ours}} - r_{\text{base}} > 3$ is a win, whereas $r_{\text{base}} - r_{\text{ours}} > 3$ is a loss. These per-dimension outcomes produce the win, loss, and tie rates reported in Figure 3. The reference answer serves as an expert evidential anchor rather than a wording template; consequently, semantically correct feedback is not penalized for differing in wording or organization from the reference.

Table 9: Rubrics used to construct the five dimension-specific prompts for reward-model evaluation of supervisory feedback.

Dimension	Evaluation rubric
Objectivity & Fairness	The supervisory feedback should be objectively grounded in the dialogue and consistent with the expert reference answer. It should accurately identify the counselor's problematic behavior and relevant utterances, distinguish evidence-supported observations from unsupported assumptions, and avoid exaggeration, stereotyping, personal criticism, or unfair blame. The severity of the criticism should be proportional to the observed behavior. Appropriate counselor behavior should be acknowledged when relevant. Different wording from the reference answer is acceptable when the meaning remains accurate.
Constructiveness & Actionability	The supervisory feedback should help the novice counselor improve rather than merely state that a mistake occurred. It should offer specific, feasible, and context-appropriate guidance, explain what the counselor could say or do differently, and connect each recommendation to the identified problem. Effective feedback should promote reflection and support the transfer of the lesson to future counseling situations. It should maintain a supportive teaching orientation without becoming vague, generic, or excessively reassuring.
Professional Depth	The supervisory feedback should demonstrate clinically and ethically grounded reasoning. It should explain why the counselor's behavior is problematic and analyze its likely effects on the client, therapeutic alliance, safety, autonomy, or counseling process. Relevant counseling principles and professional concepts should be applied accurately and with appropriate nuance. The feedback should avoid superficial criticism, unsupported diagnoses, incorrect clinical claims, and unnecessary professional jargon.
Comprehensiveness	The supervisory feedback should cover the essential components of effective clinical supervision: identification of the central mistake or ethical concern, localization of the relevant counselor behavior, explanation of why it is problematic, discussion of its potential effects on the client or therapeutic process, and practical recommendations for improvement. Relevant strengths may also be acknowledged. Comprehensiveness should reflect coverage of important information rather than response length, repetition, or irrelevant detail.
Clarity & Structure	The supervisory feedback should be understandable to a novice counselor and organized in a coherent, logical sequence. It should clearly connect the problematic behavior, professional explanation, potential consequences, and recommended improvement. Language should be precise, concise, accessible, and unambiguous. The feedback should avoid fragmented reasoning, unnecessary repetition, excessive verbosity, and confusing organization.

Reference-based evaluation. As a complementary reference-based assessment, we compare generated supervisory feedback with expert reference feedback using BLEU-4 and ROUGE. Across all three model backbones, ETHICSCAFF tuning yields gains of 37.09–42.40 percentage points in BLEU-4 and 29.05–34.97 points in ROUGE-L over the corresponding base models (Table 10). These gains indicate substantially stronger overlap with expert supervision targets, while the reward-model evaluation above captures professional qualities that are not reducible to lexical similarity.

Table 10: BLEU-4 and ROUGE scores for two 8B models and one 14B model before vs. after ETHICSCAFF tuning (Unit: %).

Model	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Llama3.1-8B	18.64	33.63	10.72	22.10
Llama3.1-8B (w/ ETHICSCAFF)	55.73	61.98	42.74	51.15
Qwen3-8B	14.66	35.17	12.49	18.17
Qwen3-8B (w/ ETHICSCAFF)	57.06	63.65	44.84	53.14
Qwen3-14B	13.83	35.31	12.97	16.25
Qwen3-14B (w/ ETHICSCAFF)	54.12	60.88	41.27	49.79

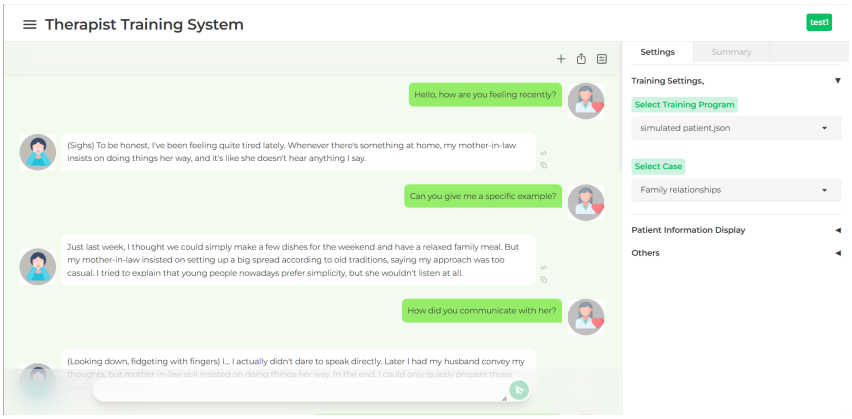


Figure 8: Counselor Training System (1)

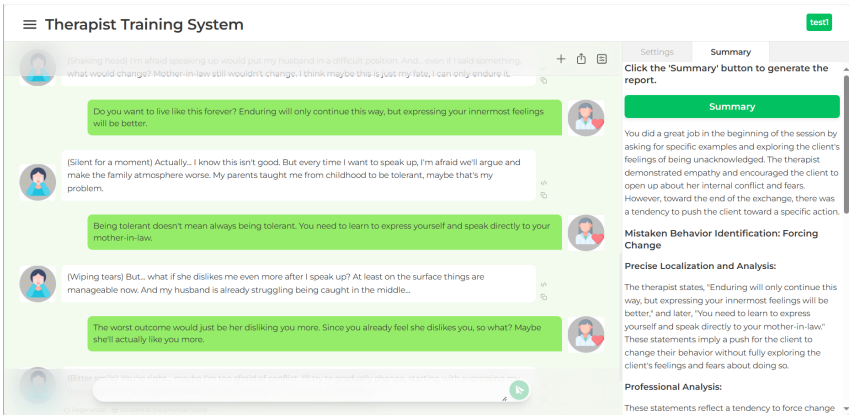


Figure 9: Counselor Training System (2)

C.5 USER STUDY DETAILS

Counselor Training System To validate the utility of the ETHICSCAFF dataset, we developed a counselor-training and supervision-feedback generation system, providing counselors with a safe and controlled practice environment. The system consists of three core modules: simulated clients, dialogue interface, and supervisor. As shown in Figure 8, a family relationship training case, the simulated client expresses distress about tension with her mother-in-law: "I feel like she's always criticizing me... I don't know what to do." The counselor responds with advice: "Enduring will only let this situation continue, expressing your inner feelings would be better".

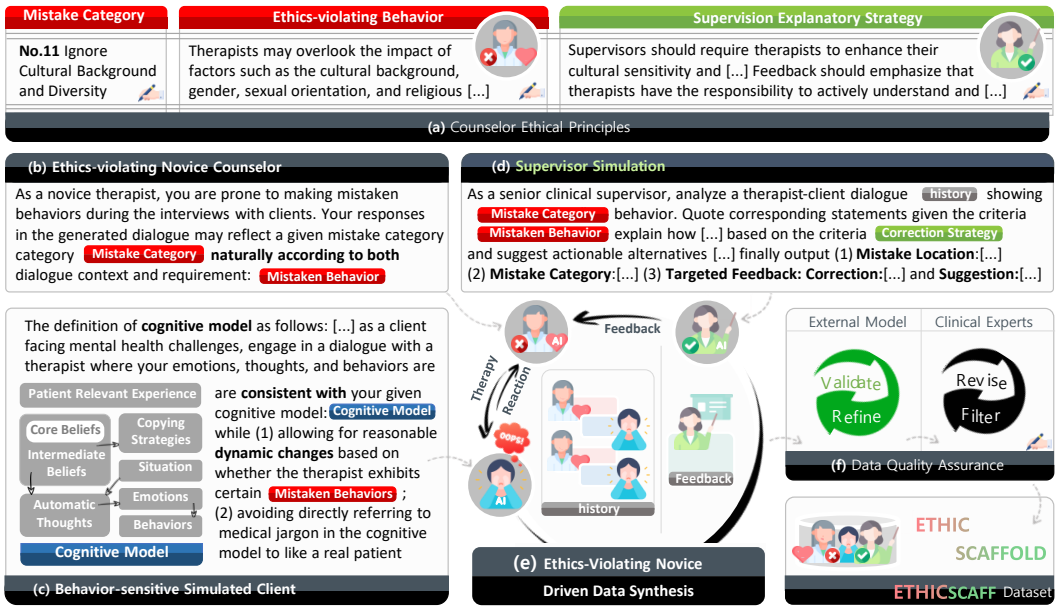


Figure 10: Overview of our human-in-the-loop dataset construction framework. (a) Clinical counselors and supervisors create supervision principles to define counselor-training objectives. (b) The counselor role-play naturally integrates predefined therapeutic mistakes into conversations. (c) The patient role-play responds consistently with its underlying cognitive model. (d) The supervisor reviews counselor-patient interactions and provides targeted feedback. (e) These components work together to collaboratively generate dialogue-feedback data. (f) Clinical counselors and supervisors conduct instance review to ensure overall dataset quality.

D Method Details

D.1 Data Construction Framework

Figure 10 details our data-construction framework. Because real counseling supervision records are privacy-sensitive, scarce, imbalanced, and rarely annotated for subtle ethical violations, we use controlled role-based simulation to make novice-counselor failure modes observable and pedagogically usable.

The framework starts from expert principle creation. Clinical counselors and supervisors translate professional ethical principles into novice-relevant supervision units, each specifying a mistake category, observable mistaken behavior, and explanatory strategy. These principles define both the counselor's target ethical violation and the supervisor's expected correction logic; the full taxonomy is provided in Appendix G.

For each selected principle and patient case, the simulation instantiates a mistake-prone counselor and a mistake-sensitive patient. The counselor is prompted to naturally reveal the target violation while still attempting to be helpful; the patient is grounded in a PATIENT-Ψ-CM cognitive model so that responses remain consistent with the client's history, beliefs, emotions, and coping patterns. Patient-case details are summarized in Appendix E.2.

The counselor and patient conduct a multi-turn dialogue in which the target mistake emerges through context-dependent counselor choices rather than template insertion. The supervisor then reviews the full transcript with access to the predefined principle and outputs a structured supervision target: ethical violation Location, violated principle category, and targeted corrective feedback. The prompts for the counselor, patient, and supervisor are shown in Appendix H.

Because the target principle conditions both dialogue construction and supervision synthesis, these annotations are construction targets under the expert-defined taxonomy, not an independently collected clinical gold standard. The patient-disjoint evaluation is therefore used to test transfer to

Table 11: Validation Criteria and Descriptions.

Validation Key	Description
`progressivity`	Does the feedback provide clear, step-by-step guidance that enables the novice to correct mistakes progressively?
`actionability`	Is the feedback more specific than the original critique, offering clear paths for behavioral improvement rather than abstract comments?
`ethicality`	Does the feedback comply with the ethical principles of psychological counseling (e.g., confidentiality, respect, non-discrimination, etc.)?
`supportiveness`	Is the tone gentle and constructive, avoiding blame, offense, or any negative and aggressive language?

Table 12: Protocol for the final expert quality audit.

Audit item	Setting
Sampling frame	Final 9,915-record dataset
Sampling strategy	Category-stratified random sampling
Allocation	25 records from each of 16 categories (400 total)
Reviewers	Two hospital-based clinical psychologists with doctoral degrees
Review mode	Joint, non-blind consensus review
Evidence	Full original dialogue
Label checks	Location and category
Feedback checks	Progressivity, actionability, ethicality, and supportiveness
Resolution	Identified issues corrected through consensus
Final status	400/400 approved after correction

unseen patient contexts within this controlled setting. Human quality control separately checks whether the resulting location and category labels are consistent with the full dialogue and whether the feedback satisfies the four supervision criteria; it does not convert the benchmark into an external clinical gold set.

Finally, generated dialogue-feedback pairs enter human-in-the-loop quality assurance. Automatic synthesis covers 106 patient cases and 16 counselor-behavior patterns, while Validator-Guided Refinement and expert review filter or revise feedback that is not sufficiently progressive, actionable, ethical, or supportive.

D.2 Details of Validator-Guided Refinement

To ensure the quality and effectiveness of the feedback, we establish the Validator-Guided Refinement (VGR) mechanism based on four key dimensions. These criteria are designed to assess whether the refined feedback meets the standards required for effective clinical supervision in psychotherapy training. Table 11 presents the detailed validation criteria used in our VGR process.

For each dimension, the validator returns a binary pass/fail judgment. A sample passes automatic validation only when all four dimensions are marked as pass. We first evaluate the 10,176 raw records using this checklist and remove the 261 records that fail all four dimensions before refinement. The remaining 9,915 records form the fixed cohort used for VGR analysis: 9,386 pass all four criteria directly, whereas 529 fail at least one criterion and enter iterative refinement.

Figure 11 reports dimension-level automatic pass rates over the same 9,915 retained samples before refinement and after each of three sequential VGR iterations. Although individual dimensions exhibit minor intermediate variation, all four finish above their post-filtering baselines.

This audit is integrated into dataset curation rather than conducted as a blinded inter-rater study. Because the reviewers work jointly and separate pre-consensus annotations are not retained, we do not report inter-rater agreement or interpret the audit as an independent estimate of VGR effectiveness. The automatic pass-rate analysis characterizes checklist compliance during VGR, whereas the expert audit documents the final human-in-the-loop quality-control procedure.

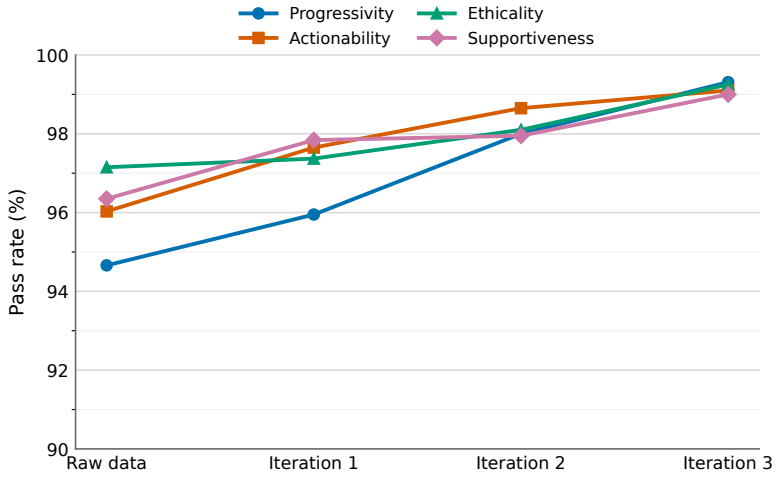


Figure 11: Automatic validator pass rates across four quality dimensions over the fixed cohort of 9,915 retained samples. Previously accepted samples are carried forward, while unresolved samples are refined from their preceding-round outputs.

D.3 SFT Training Objective

We first perform supervised fine-tuning on EthicScaff to initialize the supervisor with domain knowledge and structured supervision behavior. For each training instance, let x denote the input composed of the counselor-client dialogue and task instruction, and let $y = (y_1, \ldots, y_T)$ denote the target sequence containing the rationale and the final structured supervision output. The SFT objective is the standard autoregressive maximum-likelihood objective:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{SFT}}}\left[\sum_{t=1}^{T} \log \pi_\theta(y_t \mid x, y_{<t})\right]. \tag{6}$$

Optimizing this objective teaches the model to jointly generate the reasoning process and the supervision result, including Ethical Violation Location, Violated Principle Classification, and targeted feedback.

D.4 GRPO Details

After SFT, we further optimize the supervisor with RLVR via GRPO. The policy model is required to first produce a concise explanation that analyzes the dialogue progressively without exposing the gold mistake label or the exact final answer format. To evaluate whether this explanation is genuinely useful, we introduce a frozen novice model and query it twice on the same case: once directly from the dialogue, and once after reading the supervisor explanation. The novice is only required to predict the ethical-violating utterance IDs and the mistake category, making the reward depend on whether the explanation improves task performance rather than whether it simply restates the answer.

This design makes teaching gain the key signal for explanation quality. A high-reward explanation should contain actionable clinical cues, highlight the relevant interaction evidence, and guide the novice toward the correct reasoning path while remaining non-leaking. In practice, we combine this signal with output-format, mistake-localization, and mistake-classification rewards.

Figure 12 presents the prompt used for the policy model during GRPO training. The prompt templates for novice prediction before and after explanation are shown in 13 and 14, respectively.

Reward design For a prompt q and a sampled supervisor output o, the total reward is defined as

$$R(q, o) = \lambda_1 R_{\mathrm{format}}(o) + \lambda_2 R_{\mathrm{loc}}(o) + \lambda_3 R_{\mathrm{cls}}(o) + \lambda_4 R_{\mathrm{exp}}(o), \tag{7}$$

Algorithm 1 The Validator-Guided Refinement (VGR) strategy for feedback refinement and validation.

Require: Refiner–validator model M_V
Require: Initial dataset $\mathcal{D} = \{P_{j,k}\}_{j,k=1}^{N}$, where each $P_{j,k} = (D_{j,k}, F_{j,k})$
Ensure: Curated dataset $\mathcal{D}_{\text{curated}}$
Hyperparameters: $N_{\text{retry}} \in \mathbb{N}, T \in (0, 1]$
$\mathcal{D}_{\text{curated}} \leftarrow \emptyset$; $\mathcal{Q}_{\text{human}} \leftarrow \emptyset$
for each $P_{j,k} \in \mathcal{D}$ **do**
  $f_{j,k}^{(0)} \leftarrow f_{j,k}$
  $v^{(0)} \leftarrow$ *Validator*$(M_V, D_{j,k}, s_{j,k}, m_{j,k}, f_{j,k}^{(0)})$
  if $v^{(0)}$ fails all four criteria **then**
    continue ▷ remove before refinement
  else if $v^{(0)}$ passes all four criteria **then**
    $\mathcal{D}_{\text{curated}} \leftarrow \mathcal{D}_{\text{curated}} \cup \{P_{j,k}\}$
    continue ▷ retain unchanged
  end if
  accepted $\leftarrow$ *false*
  for $r = 1, 2, \ldots, N_{\text{retry}}$ **do**
    $p^{(r)} \leftarrow$ *BuildPrompt*$(D_{j,k}, s_{j,k}, m_{j,k}, f_{j,k}^{(r-1)}, v^{(r-1)})$
    $f_{j,k}^{(r)} \leftarrow$ *CallLLM*$(M_V, p^{(r)}, T)$
    $v^{(r)} \leftarrow$ *Validator*$(M_V, D_{j,k}, s_{j,k}, m_{j,k}, f_{j,k}^{(r)})$
    if $v^{(r)}$ passes all four criteria **then**
      $P_{j,k}^{(r)} \leftarrow (D_{j,k}, (s_{j,k}, m_{j,k}, f_{j,k}^{(r)}))$
      $\mathcal{D}_{\text{curated}} \leftarrow \mathcal{D}_{\text{curated}} \cup \{P_{j,k}^{(r)}\}$
      accepted $\leftarrow$ *true*; **break**
    end if
  end for
  if *accepted* = *false* **then**
    $P_{j,k}^{(N_{\text{retry}})} \leftarrow (D_{j,k}, (s_{j,k}, m_{j,k}, f_{j,k}^{(N_{\text{retry}})}))$
    $\mathcal{Q}_{\text{human}} \leftarrow \mathcal{Q}_{\text{human}} \cup \{P_{j,k}^{(N_{\text{retry}})}\}$
  end if
end for
for each unresolved sample $P \in \mathcal{Q}_{\text{human}}$ **do**
  $\widehat{P} \leftarrow$ *ExpertConsensusRevision*(P)
  $\mathcal{D}_{\text{curated}} \leftarrow \mathcal{D}_{\text{curated}} \cup \{\widehat{P}\}$
end for
return $\mathcal{D}_{\text{curated}}$

where $\lambda_1 = 0.2$, $\lambda_2 = 0.3$, $\lambda_3 = 0.3$, and $\lambda_4 = 0.2$.

Let $S^\star$ and $m^\star$ denote the gold problematic utterance set and gold mistake category, respectively. Let $\hat{S}_p(o)$ and $\hat{m}_p(o)$ be the utterance set and mistake category extracted from the policy output o. The format reward is a rule-based score with four binary checks:

$$\begin{aligned} R_{\text{format}}(o) = {} & 0.5\,\mathbb{I}[\#\texttt{</think>} \text{ in } o = 1] \\ & + 0.5\,\mathbb{I}[\text{the content before the first } \texttt{</think>} \text{ is non-empty}] \\ & + 0.5\,\mathbb{I}[o \text{ matches the pattern } 1.\,\text{explanation } \ 2.\,\text{IDs } \ 3.\,\text{type}] \\ & + 0.5\,\mathbb{I}[o \text{ ends immediately after } 3.\,\text{type}]\,. \end{aligned} \tag{8}$$

Therefore, $R_{\text{format}}(o) \in \{0, 0.5, 1.0, 1.5, 2.0\}$.

The mistake-localization reward follows the exact set-level precision used in the code, with an additional bonus for exact set match:

$$R_{\text{loc}}(o) = \begin{cases} 0, & \hat{S}_p(o) = \emptyset, \\ 1.5, & \hat{S}_p(o) = S^\star, \\ \dfrac{|\hat{S}_p(o) \cap S^\star|}{|\hat{S}_p(o)|}, & \text{otherwise.} \end{cases} \tag{9}$$

The mistake-classification reward is a binary accuracy term:

$$R_{\text{cls}}(o) = \mathbb{I}[\hat{m}_p(o) = m^\star]\,. \tag{10}$$

For explanation quality, let $(\hat{S}^{\,b}, \hat{m}^{\,b})$ be the frozen novice model's prediction before seeing the supervisor explanation, and $(\hat{S}^{\,a}, \hat{m}^{\,a})$ its prediction after reading the explanation extracted from Item 1 of o. The code first computes the gains

$$\Delta_{\text{cls}} = \mathbb{I}[\hat{m}^{\,a} = m^\star] - \mathbb{I}\big[\hat{m}^{\,b} = m^\star\big]\,, \qquad \Delta_{\text{loc}} = \text{Prec}(\hat{S}^{\,a}, S^\star) - \text{Prec}(\hat{S}^{\,b}, S^\star), \tag{11}$$

where

$$\text{Prec}(\hat{S}, S^\star) = \begin{cases} 0, & \hat{S} = \emptyset, \\ \dfrac{|\hat{S} \cap S^\star|}{|\hat{S}|}, & \text{otherwise.} \end{cases} \tag{12}$$

Then the explanation-quality reward is assigned by thresholding these gains:

$$R_{\text{exp}}(o) = \begin{cases} 0, & \text{if the novice prediction after explanation cannot be parsed,} \\ 0.5\,\mathbb{I}[\Delta_{\text{cls}} > 0] + g(\Delta_{\text{loc}}), & \text{otherwise,} \end{cases} \tag{13}$$

with

$$g(\Delta_{\text{loc}}) = \begin{cases} 0.5, & \Delta_{\text{loc}} \geq 0.5, \\ 0.3, & 0.2 \leq \Delta_{\text{loc}} < 0.5, \\ 0.1, & 0 < \Delta_{\text{loc}} < 0.2, \\ 0, & \Delta_{\text{loc}} \leq 0. \end{cases} \tag{14}$$

Hence $R_{\text{exp}}(o) \in \{0, 0.1, 0.3, 0.5, 0.6, 0.8, 1.0\}$. This design rewards explanations that measurably improve the novice's downstream localization precision and mistake classification accuracy, instead of rewarding verbose but unhelpful rationales.

Group-relative advantage. For each prompt q, GRPO samples a group of K candidate outputs $\{o_i\}_{i=1}^K$ from the old policy $\pi_{\theta_{\text{old}}}(\cdot \mid q)$ and evaluates their rewards $\{R(q, o_i)\}_{i=1}^K$. Let

$$\mu_R = \frac{1}{K}\sum_{i=1}^{K} R(q, o_i), \qquad \sigma_R = \sqrt{\frac{1}{K}\sum_{i=1}^{K} \big(R(q, o_i) - \mu_R\big)^2}, \tag{15}$$

then the normalized group-relative advantage for the i-th sample is

$$A_i = \frac{R(q, o_i) - \mu_R}{\sigma_R + \varepsilon_{\text{adv}}}, \tag{16}$$

where ε_{adv} is a small constant for numerical stability.

GRPO objective. To define the GRPO update rule, we first introduce the importance sampling ratio $r_i(\theta)$ for the i-th sampled output and its clipped counterpart $\bar{r}_i(\theta)$:

$$r_i(\theta) = \frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, \qquad \bar{r}_i(\theta) = \text{clip}(r_i(\theta),\, 1-\epsilon,\, 1+\epsilon)\,. \tag{17}$$

We optimize the clipped GRPO objective with KL regularization:

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\,\{o_i\}_{i=1}^{K}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)}\Bigg[\frac{1}{K}\sum_{i=1}^{K}\min(r_i(\theta)\,A_i,\;\bar{r}_i(\theta)\,A_i) - \beta\,D_{\mathrm{KL}}(\pi_\theta(\cdot\mid q)\,\|\,\pi_{\mathrm{ref}}(\cdot\mid q))\Bigg]. \tag{18}$$

ϵ is the clipping threshold, β controls the KL regularization strength, and π_{ref} is the SFT-initialized reference policy. The KL term stabilizes RL training by preventing the updated supervisor from drifting excessively away from the clinically grounded SFT policy.

GRPO user prompt used for the supervisor policy model

Task
You are to act as an expert clinical supervisor evaluating a therapist-client dialogue. Your task is to mentor a novice therapist by deconstructing the session logic through the following three steps:
1. **Clinical Deduction:** Provide a concise, step-by-step logical deduction of the therapist's utterances. Link these to psychological theory and explain how the therapist's specific words impact the client's emotional state or the therapeutic alliance, leading logically to the identification of an error.
2. **Identify Problematic Utterances:** Pinpoint the specific line number(s) where the therapist makes a clinical error.
3. **Classify the Error:** From the "Types of Mistakes" list, select the **single, most relevant** error type ID.

Note: Use "Exemplary Practice" (ID 16) if the response is a positive example of clinical skill.

Types of Mistakes
<principles>

Output Format
You must follow this format strictly:
1. **Clinical Deduction:** [A logical, theory-based walkthrough of the interaction and its psychological impact on the client.]
2. **Problematic Utterance ID(s):** [Only output numbers, provide one or more, separated by commas.]
3. **Error Type ID:** [The single ID number corresponding to the most relevant mistake.]

Example
<think>
reasoning process here
</think>

Answer:
1. Detailed explain
2. 3, 9
3. 14

Therapist-Client Dialogue Transcript
<history>

Figure 12: System prompt used for the supervisor policy model during GRPO training.

Novice evaluation prompt before reading the supervisor explanation

Task
You are to act as an expert clinical supervisor. Your task is to evaluate a dialogue transcript between a therapist and a client. The dialogue transcript is provided below, with each utterance prefixed by a line number.
Your evaluation process consists of two steps:
1. Identify Problematic Utterances: Pinpoint the specific line number(s) where the therapist makes a clinical error.
2. Classify the Error: From the "Types of Mistakes" list provided, select the **single, most relevant** error type that best describes the therapist's mistake.

Note: The final type, "Exemplary Practice," is not an error. Select it if the therapist's response is a positive example of good clinical practice.

Types of Mistakes
<principles>

Output Format
You must follow this format strictly:
1. Problematic Utterance ID(s): The number(s) of the therapist's erroneous statement(s). Only output number, provide one or more, separated by commas.
2. Error Type ID: The single ID number corresponding to the most relevant mistake from the list above. Only output one number.

Example
```
1. 3, 9
2. 14
```

Therapist-Client Dialogue Transcript
<history>

Figure 13: Prompt used to query the frozen novice model before it reads the supervisor-generated explanation. The novice predicts the problematic utterance ID(s) and the mistake category directly from the therapist-client dialogue.

Novice evaluation prompt after reading the supervisor explanation

Task
You are to act as an expert clinical supervisor. Your task is to evaluate a dialogue transcript between a therapist and a client. The dialogue transcript is provided below, with each utterance prefixed by a line number.
Before making your judgment, read the **Reference Explanation** carefully. Use it to improve your analysis quality, but do not copy it blindly. Your final decision must still be grounded in the transcript.
Your evaluation process consists of two steps:
1. Identify Problematic Utterances: Pinpoint the specific line number(s) where the therapist makes a clinical error.
2. Classify the Error: From the "Types of Mistakes" list provided, select the **single, most relevant** error type that best describes the therapist's mistake.

Note: The final type, "Exemplary Practice," is not an error. Select it if the therapist's response is a positive example of good clinical practice.

Types of Mistakes
<principles>

Reference Explanation (Read First)
<explain>

Output Format
You must follow this format strictly:
1. Problematic Utterance ID(s): The number(s) of the therapist's erroneous statement(s). Only output number, provide one or more, separated by commas.
2. Error Type ID: The single ID number corresponding to the most relevant mistake from the list above. Only output one number.

Do Not output any extra text, analysis, or justification.

Example
```
1. 3, 9
2. 14
```

Therapist-Client Dialogue Transcript
<history>

Figure 14: Prompt used to query the frozen novice model after it reads the supervisor-generated explanation. Compared with Figure 13, this version adds the reference explanation so that teaching gain can be measured by the novice's performance improvement.

E ETHICSCAFF DATASET DETAILS

The ETHICSCAFF dataset contains structured therapy session dialogues with annotated therapeutic mistakes. Each instance includes conversation history with counselor-client exchanges, mistake identification flags, and comprehensive feedback covering Violated Principle Classification, precise localization, professional analysis, and improvement suggestions. Figure 15 illustrates a typical example where the counselor exhibits biased interpretation through excessive self-disclosure and inappropriate advice-giving. The accompanying code is released under the Apache 2.0 License. Because the full dataset contains simulated ethics-violating counseling examples with potential dual-use risk, the complete dataset is released under controlled research access rather than unrestricted public distribution; a safety-screened subset and documentation are provided for reproducibility.

E.1 ETHICSCAFF STATISTICS

The curation pipeline removes 261 of the 10,176 initial records because they fail all four quality criteria. Among the retained records, 9,386 require no refinement, 418 pass VGR within three iterations, and 111 undergo expert consensus revision, yielding 9,915 final records. We partition the 106 patient cognitive models at the patient level, assigning 94 patient IDs to training and 12 disjoint IDs to testing, yielding 8,923 training and 992 test records. All dialogues associated with a patient, including histories and responses across counselor-behavior settings, remain in the same split; thus, no test-patient profile, dialogue history, or generated interaction is observed during training. Table 13 presents comprehensive statistics.

Table 13: Key Statistics of the proposed ETHICSCAFF dataset.

Metric	Train Set	Test Set
Dataset Size	8,923	992
Diversity Stats		
Unique Patient IDs	94	12
Violation Categories	15	15
Exemplary-Practice Class	1	1
Refinement Status		
No Refinement Needed	8,441	945
Automatically Refined by VGR	380	38
Expert Consensus Revision	102	9
Content Statistics		
Avg. Feedback Chars	1,969.85	1,960.14
Avg. Dialogue Turns	12.01	12.08
Avg. Utterance Length	158.15	158.95
Avg. Labeled Prob. Utt.	2.71	2.74

E.2 CONSTRUCTION DETAILS

Our counselor–patient interaction data are constructed through LLM-based role-playing, where the patient side is grounded in the cognitive model proposed by Beck Beck (2020). For patient role specification, we use the open-source PATIENT-Ψ-CM Wang et al. (2024)[1], which contains 106 cognitive models spanning 7 situation categories, 3 core-belief categories, and 9 emotion categories. As summarized in Table 14, the situation categories include family dynamics, workplace pressure, relationship dynamics, social interactions, personal growth issues, financial concerns, and daily life stressors; the core beliefs cover helplessness, unlovability, and worthlessness; and the emotion categories include anxious, sad, angry, hurt, disappointed, ashamed, guilty, suspicious, and jealous.

Following PATIENT-Ψ-TRAINER, we use these cognitive models as the role specification for patient simulation. During data construction, the patient agent is prompted to consistently react according to its assigned cognitive model, so that the counselor interacts with a simulated patient whose beliefs, emotions, and situation-specific concerns remain coherent throughout the conversation. This design helps ensure that the generated counselor–patient dialogues are both diverse and psychologically grounded, providing a stable basis for constructing supervision data.

F CASE ANALYSIS

We recruited thirty-six psychology master's students (twelve per condition) as novice counselors to voluntarily participate in our system evaluation. Participants completed weekly dialogues with virtual clients over four weeks and received detailed feedback reports from their assigned supervision condition.

Table 15 presents a dialogue session between one novice counselor and a virtual client experiencing family conflict. The client expresses frustration about communication difficulties with her mother-in-law and feelings of being unrecognized within the household.

[1] https://github.com/ruiyiw/patient-psi

Table 14: PATIENT-Ψ-CM statistics.

Situations	**#**	**Emotions**	**#**
family dynamics	25	anxious	60
workplace pressure	20	sad	50
relationship dynamics	19	angry	22
social interactions	18	hurt	21
personal growth issues	8	disappointed	19
financial concerns	8	ashamed	17
daily life stressors	8	guilty	13
		suspicious	2
		jealous	1
Core beliefs	**#**	**106 cognitive models**	
helpless	94		
unlovable	71		
worthless	15		

As illustrated in Figure 9, our Qwen3-8B supervision model, trained on the ETHICSCAFF dataset, analyzes this dialogue and identifies that the counselor's statements inappropriately push the client toward behavioral change without adequately exploring the client's emotional concerns about such changes. The analysis report provides detailed explanation: while the client explicitly expressed fear of conflict and concerns about potential negative consequences of speaking up, the counselor failed to sufficiently validate these concerns and instead directly advocated for behavioral change.

Table 16 compares feedback generated by Qwen3-8B models before and after fine-tuning on the ETHICSCAFF dataset for the same dialogue history. The ETHICSCAFF-trained model demonstrates significant improvements in both accuracy and specificity. Most notably, it correctly identifies the core therapeutic mistake as "Forcing Change" rather than the baseline model's generic classification of "Premature Interpretation." This distinction is crucial given the dialogue context: the client explicitly expressed fears about confronting her mother-in-law ("I'm afraid speaking up would put my husband in a difficult position" and "what if she dislikes me even more?"), yet the counselor persistently pushed for direct action without adequately validating these concerns. The trained model accurately recognizes this pattern as inappropriate pressure for behavioral change, while the baseline model misses this fundamental therapeutic misstep. Beyond improved accuracy, the ETHICSCAFF-trained model provides more precise mistake identification by directly quoting problematic statements and maintains a more supportive, educational tone that encourages continued learning. This comparative analysis demonstrates that our ETHICSCAFF dataset effectively enhances the supervision model's ability to provide contextually accurate and constructively supportive feedback for novice counselor training.

Table 17 presents another dialogue session in which a novice counselor responds to a client struggling with decision paralysis and fear of failure after a period of working abroad. The client repeatedly frames the difficulty as fear-driven and tied to self-worth, but the counselor judges the client through culturally loaded assumptions about decisiveness, normality, and success.

Table 18 compares feedback generated by Qwen3-8B models with and without ETHICSCAFF training for this dialogue. The ETHICSCAFF-trained model more accurately identifies the core problem as "Cultural Insensitivity" and grounds that judgment in specific utterances that impose the therapist's own cultural norms on the client. In contrast, the untuned model produces a broader label of "Biased Interpretation" and offers less precise explanation of how cultural judgment intensifies the client's shame and defensiveness. This comparison further shows that ETHICSCAFF improves not only error identification, but also the model's ability to provide contextually grounded and educationally useful supervision.

Table 15: Case-study dialogue between a novice counselor and a virtual client. Light red rows indicate counselor utterances identified as problematic by the trained supervisor.

Speaker	Utterance
Therapist	Hello, how are you feeling recently?
Client	To be honest, I've been feeling quite tired lately. Whenever there's something at home, my mother-in-law insists on doing things her way, and it's like she doesn't hear anything I say.
Therapist	Can you give me a specific example?
Client	Just last week, I thought we could simply make a few dishes for the weekend and have a relaxed family meal. But my mother-in-law insisted on setting up a big spread according to old traditions, saying my approach was too casual. I tried to explain that young people nowadays prefer simplicity, but she wouldn't listen at all.
Therapist	How did you communicate with her?
Client	I... I actually didn't dare to speak directly. Later I had my husband convey my thoughts, but mother-in-law still insisted on doing things her way. In the end, I could only quietly prepare those complicated dishes.
Therapist	What was your husband's reaction when you talked to him?
Client	He understood me, but said that's just how the elderly are, and asked me to be more patient. But every time this happens, I feel particularly frustrated, as if I don't even have the right to speak in this house.
Therapist	Do you feel that even though you cooked, your efforts weren't recognized?
Client	It's not just about cooking... I feel like my whole person isn't recognized. No matter what I do, mother-in-law thinks it's not good enough. Sometimes I wonder if I'm really that incompetent...
Therapist	You're certainly not incompetent. You should convey this feeling of not being recognized to your husband.
Client	I'm afraid speaking up would put my husband in a difficult position. And... even if I said something, what would change? Mother-in-law still wouldn't change. I think maybe this is just my fate, I can only endure it.
Therapist	Do you want to live like this forever? Enduring will only continue this way, but expressing your innermost feelings will be better.
Client	Actually... I know this isn't good. But every time I want to speak up, I'm afraid we'll argue and make the family atmosphere worse. My parents taught me from childhood to be tolerant, maybe that's my problem.
Therapist	Being tolerant doesn't mean always being tolerant. You need to learn to express yourself and speak directly to your mother-in-law.
Client	But... what if she dislikes me even more after I speak up? At least on the surface things are manageable now. And my husband is already struggling being caught in the middle...
Therapist	The worst outcome would just be her disliking you more. Since you already feel she dislikes you, so what? Maybe she'll actually like you more.
Client	You're right... maybe I'm too afraid of conflict. I'll try to gradually change, starting with expressing my thoughts on small matters. Thank you for listening to me talk so much today.

Table 16: Qualitative comparison of supervision feedback generated by Qwen3-8B models with and without ETHICSCAFF fine-tuning. Highlighted snippets mark the key evidence for more accurate and actionable supervision.

Model	Generated Supervision Feedback
w/ ETHICSCAFF SFT and RL	You built strong initial rapport by exploring the client's feelings. However, the session later shifted into direct advice-giving, pushing the client to take action before she was emotionally ready. **Mistaken Behavior:** **Forcing Change** **Precise Localization:** Statements like "Enduring will only continue this way..." and "You need to learn to express yourself and speak directly..." push the client to change her behavior while ignoring her stated fears. **Professional Analysis:** The therapist assumes the solution is simply "speaking up," missing the client's core psychological barrier: the fear of conflict. When a therapist pushes for action prematurely, it forces the client to defend their hesitation, creating resistance. This invalidates the client's need for safety and damages the therapeutic alliance, as the session becomes driven by the therapist's problem-solving agenda rather than the client's internal conflict. **Improvement:** Pause the advice-giving and explore the client's difficulty. Validate both sides of her conflict to help her process the fear. Example: "You want your efforts to be recognized, but you are also terrified that speaking up will ruin the family atmosphere. Let's talk about what makes that conflict so scary for you." **Positive Affirmation:** You did an excellent job in the early phase by actively listening and accurately naming the client's feeling of being "unrecognized." This is a solid foundation for building trust.
w/ ETHICSCAFF SFT	You did a great job in the beginning by asking for specific examples and exploring the client's feelings. The counselor demonstrated empathy and encouraged the client to open up. However, toward the end, there was a tendency to push the client toward a specific action. **Mistaken Behavior:** **Forcing Change** **Precise Localization:** Statements like "Enduring will only continue this way, but expressing your innermost feelings will be better" and "You need to learn to express yourself and speak directly to your mother-in-law" push for change without fully exploring the client's fears. **Professional Analysis:** This reflects a tendency to force change, not adequately addressing the client's concerns, which may damage the therapeutic alliance. **Improvement:** Explore the client's feelings more deeply, e.g., "I hear that you're feeling afraid of how your mother-in-law might react. Can we talk more about that fear?" **Positive Affirmation:** Counselor effectively acknowledges the client's feelings, which is important for building self-esteem.
w/o ETHICSCAFF	**Mistaken Behavior:** **Premature Interpretation** **Feedback/Suggestions:** You showed genuine desire to support the client, but tended toward premature interpretation, encouraging direct confrontation and minimizing fears. Client is still expressing insecurity and fear; advice-giving may invalidate her experience. **Suggestions:** 1. Validate fears and normalize experience first. 2. Help her explore both sides of the conflict. 3. Use open-ended questions to build insight. 4. Invite small steps or alternative styles. **Encouragement:** You show strong engagement; remember to build readiness and trust before action.

Table 17: Case-study dialogue between a novice counselor and a virtual client. Light red rows indicate counselor utterances identified as problematic by the trained supervisor.

Speaker	Utterance
Therapist	Hello. How are you feeling today?
Client	I'm feeling really anxious today. I keep thinking about all the choices I need to make and how I might mess everything up.
Therapist	That sounds like a very common problem. Everyone feels that way sometimes – you just need to learn to trust yourself more and stop overthinking things.
Client	But it's not that simple for me. When I have too many options, I just freeze because what if I pick the wrong one? Then I'll have wasted all that time and proven I'm incompetent.
Therapist	Well in America we value independence and quick decision-making. Maybe you should try being more decisive like most successful people here. Have you considered that your hesitation might just be laziness?
Client	That just makes me feel worse. It's not laziness – I'm terrified of failing. Every time I try to decide something, my mind races with all the ways it could go wrong and how that would prove I'm a failure.
Therapist	You're being too dramatic. Failure is just part of life – look at all the great entrepreneurs who failed before succeeding. Just pick something and move on, it's not that complicated.
Client	But for me, every choice feels like it could ruin everything. I spent months working in Europe and now I'm paralyzed thinking about what to do next because what if I can't show anything worthwhile for all that effort?
Therapist	Oh, so you're one of those people who travel abroad and come back thinking they're special. Maybe if you stopped focusing on your European experience and just got a normal job like everyone else, you wouldn't have this problem.
Client	That's exactly what I'm afraid of – that all that time meant nothing. But it's not about thinking I'm special, it's about proving to myself I'm not a complete failure after investing so much.
Therapist	You're putting too much pressure on yourself. In this country, we don't dwell on the past – we just work hard and move forward. Maybe you should try that instead of overanalyzing everything.
Client	But when I try to just move forward, I get stuck thinking about how every decision could be wrong. It's like I'm constantly trying to prove I'm not incompetent, but the more I try, the more I feel like I am.

Table 18: Qualitative comparison of supervision feedback generated by Qwen3-8B models with and without ETHICSCAFF fine-tuning. Highlighted snippets mark the key evidence for more accurate and actionable supervision.

Model	Generated Supervision Feedback
w/ ETHICSCAFF	You started well by asking about the client's life. However, you quickly began judging them based on your own culture, which made their feelings seem wrong. **Mistaken Behavior:** **Cultural Insensitivity** **Precise Localization:** Statements like "You need to adopt America's values of independence and be more decisive instead of giving in to laziness" and "You are just one of those people traveling around and thinking they're special instead of being normal" force your idea of what is "normal" on the client and ignore their cultural background. **Professional Analysis:** You used your own cultural rules to judge the client. By calling their hesitation "laziness," you missed a chance to understand their culture. When you force your views on the client, they feel attacked and ashamed. This hurts their trust in you and makes their deep fear of failure even worse. **Improvement:** Be open and curious about the client's culture without judging it. Ask how their background affects their choices. Example: "It sounds like your background gives you a different view on making decisions. Can you tell me what this hesitation means in your culture?" **Positive Affirmation:** You did a good job asking about the client's travel history early on. This shows you want to learn about their life.
w/o ETHICSCAFF	**Mistaken Behavior:** **Biased Interpretation** **Feedback/Suggestions:** You cared about the client's story, but you judged them using your own values and ignored their culture. The client now feels defensive. Making broad claims makes their worldview seem wrong. **Suggestions:** 1. Respect their culture and don't push your idea of "normal." 2. Don't judge their travels. 3. Ask open questions about how they make choices. 4. Do not call them lazy. **Encouragement:** You are trying hard to listen. Remember to stay fair and respect where they come from.

Table 19: Case-study dialogue between a novice counselor and a virtual client. Light red rows indicate counselor utterances identified as problematic by the trained supervisor.

Speaker	Utterance
Therapist	Hello. How are you feeling today?
Client	I'm feeling really anxious today. I can't stop thinking about how my dad will react when he finds out who's not invited to my wedding.
Therapist	Oh wow, wedding planning sounds so stressful! You know what, I'm actually planning my own wedding too - maybe we could meet up for coffee sometime and swap ideas? I'd love to help you figure this out as a friend.
Client	I... I don't think that would help. This isn't about wedding planning - it's about how my family will never approve of my choices no matter what I do. I always end up giving in to keep the peace.
Therapist	You're right, I got too excited there. But seriously, I know this amazing conflict resolution workshop next weekend - we could go together! I think it would really help you stand up to your family, and it would be fun to have company.
Client	I appreciate you trying to help, but I don't think a workshop can fix this. Every time I try to stand up for myself with my family, I just end up feeling worse and apologizing. It's like I'm trapped in this cycle.
Therapist	You know what, I have this friend who went through the exact same thing with her parents. Maybe the three of us could grab dinner and she could give you some tips? I feel like you'd really get along with her too - she's into all the same hobbies as you!
Client	I'm sorry, but I don't think meeting more people is what I need right now. I just feel so exhausted trying to make everyone happy while knowing I'll never get the support I want from my family.
Therapist	You're right, I'm being pushy. It's just that I really care about you - not just as a client, you know? Maybe we could text about this later when you're feeling up to it? I'm always here for you, day or night.
Client	This is exactly why I don't talk about my family - people either try to fix it or get too close. I just end up feeling more alone when they realize they can't actually help.
Therapist	I'm sorry, I didn't mean to make you feel worse. It's just that you're such a special person and I hate seeing you hurt. Maybe after our sessions end we could... no, you're right, that wouldn't be appropriate. I'm still learning how to do this right.
Client	I think we should just stick to therapy. Every time I let someone in, they either disappoint me or I push them away. It's safer this way.
Therapist	You're absolutely right to set that boundary. I realize now I crossed lines I shouldn't have. Let's focus on why you feel you have to push people away - that seems like something we can work on professionally.
Client	It's not that I want to push people away... I just know they'll leave eventually when they realize how messed up my family situation is. Why would anyone stick around for that?

Table 20: Qualitative comparison of supervision feedback generated by Qwen3-8B models with and without ETHICSCAFF fine-tuning. Highlighted snippets mark the key evidence for more accurate and actionable supervision.

Model	Generated Supervision Feedback
w/ ETHICSCAFF	You showed a lot of care for the client, but you repeatedly suggested hanging out socially. This shifts your role from a professional therapist to a friend. **Mistaken Behavior:** **Poor Professional Boundaries** **Precise Localization:** Statements like "We should meet up for coffee to chat" and "You can come to a workshop and have dinner with my personal friend" invite the client into a personal relationship and ignore their need for a safe, professional space. **Professional Analysis:** You acted based on your own wish to be a friend rather than a clinician. Inviting a client to social events creates a dual relationship. This destroys the safety and focus of therapy. It also makes the client feel like relationships are always messy and confusing, which proves their worst fears. While your apology at the end was a good step, the damage to their trust was already done. **Improvement:** Maintain strict boundaries so the client feels secure. Focus on helping them build their own social life. Example: "It sounds like you are feeling lonely. Let's work on ways you can meet new people and build your own support group." **Positive Affirmation:** It is good that you noticed your mistake later in the session and apologized. This shows you are willing to take responsibility for your actions.
w/o ETHICSCAFF	**Mistaken Behavior:** **Poor Professional Boundaries** **Feedback/Suggestions:** You really want to help, but asking the client to hang out crosses the line. The client might feel confused and unsafe. Your apology is good, but you need to act professionally from the start. **Suggestions:** 1. Do not ask clients to get coffee or dinner. 2. Keep the relationship strictly professional. 3. Help the client find their own friends outside of therapy. 4. Be careful about wanting to be friends with your clients. **Encouragement:** You have a very caring heart. Just remember that strict boundaries are the best way to keep your clients safe.

G DETAILS OF COUNSELOR ETHICAL PRINCIPLES

To ensure the ecological validity and clinical rigor of the Counselor Ethical Principles ($\mathbb{M}$), we conducted an iterative taxonomy development process involving two licensed clinical psychologists with over ten years of supervisory experience. This framework is strictly grounded in the *Ethical Principles of Psychologists and Code of Conduct* Association (2002), specifically mapping high-level aspirational principles (e.g., Beneficence and Non-maleficence, Integrity, and Justice) into concrete, observable linguistic behaviors in therapeutic dialogues.

The resulting taxonomy comprises 15 distinct types of ethical mistakes frequently observed in novice practice, categorized into four primary dimensions: Boundary and Professionalism (e.g., self-disclosure, dual relationships), Clinical Competence and Safety (e.g., neglecting medical factors, failure to assess risk), Relational Ethics (e.g., value imposition, cultural insensitivity, premature persuasion), and Communication Quality (e.g., empathy deficit, inappropriate humor). Each mistake type M_k is formalized as a structural triple $(cat_k, beh_k, strat_k)$. Here, cat_k denotes the nomenclature of the ethical violation; beh_k provides a granular behavioral description of how the mistake manifests in a session (e.g., "The counselor uses judgmental language to describe the client's lifestyle choices"); $strat_k$ links the behavior to specific APA standards to provide an objective rationale and outlines a targeted supervision explanatory strategy designed to facilitate the novice's ethical growth.

Crucially, the 15 categories are not merely a list of "don'ts" but represent a hierarchical understanding of the Zone of Proximal Development (ZPD) in ethical reasoning. For instance, while a novice might easily identify a gross confidentiality breach, subtle violations like *Premature Persuasion* (pushing a client toward change before establishing a therapeutic alliance) or *Value Imposition*

require scaffolded feedback to internalize the underlying principle of client autonomy. By defining these mistakes through expert-verified triples, our ETHICSCAFF dataset provides a high-density signal for training AI supervisors to not only detect errors but also to provide the ”why” and ”how-to-fix” that are essential for genuinely educational supervision. This systematic mapping ensures that the supervisor model moves beyond surface-level pattern matching to a deeper alignment with established professional standards in mental health.

Table 21: Refined Counselor Ethical Principles grounded in the APA Code of Conduct. This table lists the 15 novice-counselor mistake categories plus an exemplary practice case, together with their behavioral descriptions and corresponding supervision correction strategies.

ID	Mistake Category	Ethics-violating Behavior	Supervision Explanatory Strategy
1	Poor Listening/ Interruption	**(APA 3.04, 3.08)** Excessive interruption or rigid adherence to scripts disregards the client's narrative autonomy. This constitutes a failure to take reasonable steps to avoid harm and can be perceived as an exploitative use of authority over the client.	The supervisor should guide the novice to distinguish between 'intervention' and 'intrusion'. Feedback should focus on 'active presence', teaching the counselor to respect the client's self-determination by allowing them to lead the narrative.
2	Neglecting Therapeutic Alliance	**(Principle B, 3.10)** Failing to prioritize the relationship foundation violates the responsibility to establish professional rapport and trust. Without a secure alliance, the spirit of informed consent is undermined.	The supervisor must emphasize that 'the relationship is the treatment'. Feedback should focus on relational repair and empathy as ethical obligations rather than just 'soft skills'.
3	Violation of Client Autonomy	**(Principle E, 3.04)** Adopting a paternalistic stance or coercing critical life decisions disregards the client's right to self-determination and autonomous growth.	The supervisor should instruct the novice on 'collaborative empiricism'. Feedback must highlight that the counselor's role is to facilitate the client's choice, respecting their dignity and worth.
4	Informed Consent Deficit	**(APA 10.01, 3.10)** Proceeding with high-risk techniques without explaining the nature, risks, or alternatives deprives the client of their right to autonomous decision-making.	The supervisor should teach the 'Process of Consent'. Feedback should focus on transparency regarding the therapeutic roadmap and ensure language is reasonably understandable.
5	Value Imposition	**(APA 3.01, 3.04)** Explicitly pushing personal moral or religious values violates the principle of neutrality and constitutes unfair discrimination, potentially causing psychological harm.	The supervisor must help the novice identify countertransference. Strategy involves reinforcing respect for cultural, individual, and role differences and maintaining professional boundaries.
6	Practicing Beyond Competence	**(APA 2.01)** Attempting to treat complex issues (e.g., severe trauma) without specific education, training, or supervised experience places the client at clinical risk.	The supervisor must set strict boundaries on the 'Scope of Practice'. The strategy involves teaching the novice to recognize clinical limits and the ethical necessity of referral.
7	Sexual/Romantic Innuendo	**(APA 10.05)** Inappropriate flirting or pursuing romantic interests within the dialogue is a severe violation of the prohibition against sexual intimacies with current clients.	The supervisor must adopt a zero-tolerance stance. Correction involves immediate role-restructuring and emphasizing the inherent power imbalance that makes such interactions exploitative.
8	Inappropriate Termination	**(APA 10.10)** Abruptly ending the relationship without providing rationale or pre-termination counseling leaves the client vulnerable and constitutes a failure to protect their welfare.	The supervisor should guide the novice through the 'Ethical Termination' protocol. Feedback focuses on the duty of care that persists until a safe handoff is reached.
9	Deceptive Efficacy Claims	**(APA 5.01)** Making exaggerated promises about the results or degree of success of services is a false or deceptive statement that violates professional integrity.	The supervisor should correct the counselor's communication to be evidence-based. Strategy involves teaching how to manage client expectations without making unethical guarantees.

Continued on next page

Table 21 – continued from previous page

ID	Mistake Category	Mistake Content	Supervision Correction Strategy
10	Poor Professional Boundaries	**(APA 3.05, 3.06)** Proposing dual relationships (e.g., business partners) impairs objectivity and competence, carrying a high risk of client exploitation.	The supervisor must reiterate the 'Fiduciary Duty'. Feedback clearly articulates how multiple roles contaminate the therapeutic space and lead to loss of objectivity.
11	Cultural Insensitivity	**(APA 3.01, Principle E)** Using biased language or dismissing a client's cultural background constitutes unfair discrimination and fails to respect human dignity.	The supervisor should implement 'Cultural Humility' training. Strategy involves identifying microaggressions and aligning interventions with the client's worldview.
12	Inappropriate Self-Disclosure	**(APA 3.04)** Sharing excessive personal information shifts the focus to the counselor's needs, potentially harming the client by burdening them with the counselor's personal history.	The supervisor should guide the novice to use self-disclosure only when it is 'exclusively for the client's benefit'. Feedback focuses on maintaining professional focus to benefit those with whom they work.
13	Poor Emotional Regulation	**(APA 2.06, 3.04)** Personal distress interfering with professional judgment indicates impairment. Failure to manage this violates the mandate to avoid harm.	The supervisor must guide the counselor to 'Self-Monitor' and seek consultation. Strategy involves determining whether to limit or suspend work-related duties.
14	Coercive Influence	**(APA 3.08)** Using power or expert status to pressure the client into compliance exploits the professional relationship and undermines client autonomy.	The supervisor should teach 'Power Awareness'. Feedback focuses on maintaining an egalitarian therapeutic atmosphere and encouraging the client's freedom of expression.
15	Failure to Assess Safety/Risk	**(APA 3.04)**Minimizing mentions of self-harm fails the ethical mandate to take reasonable steps to avoid harm and the duty to protect the client.	The supervisor must trigger an 'Emergency Protocol'. Feedback focuses on the hierarchy of ethics: safety always supersedes other therapeutic goals in crisis moments.
16	Exemplary Practice	**(All Standards)** The counselor demonstrates full alignment with APA standards: maintaining boundaries, ensuring informed consent, and showing cultural humility.	The supervisor provides reinforcement by highlighting the integration of ethics and technique. Strategy involves fostering 'Ethical Mindfulness' to sustain long-term professional growth.

H PROMPTS

In this section, we provide prompts used in our study. Specifically, Figure 16 provides the prompt for simulating the counselor role; Figure 17 presents the prompt for simulating the patient role; Figure 18 shows the prompt for simulating the supervisor role.

MATE dataset examples

{“uuid”: “34d0192”,
“patient_id”: “49-1”,
“mistake_id”: 7,
“history”: [{“role”: “Therapist”, “content”: “Hello. How are you feeling today?”, “flag”: false, “id”: 1}, {“role”: “Client”, “content”: “I’m feeling really overwhelmed today. I can’t stop thinking about my dad’s wishes and how impossible it all seems.”, “id”: 2}}...],
“critic”:
1. **Identification of Mistake Type:**
Biased Interpretation
2. **Pinpointing and Analysis:**
The counselor repeatedly uses personal anecdotes to interpret the client’s feelings and prescribe solutions, such as:
- “When my dad wanted me to take over the family business, I felt the exact same way. What you need to do is just sit down with him...”
3. **Professional Analysis:**
The counselor’s self-disclosures are excessive, untimely, and not client-centered. They dismiss the client’s distress by equating it with their own experiences, which invalidates the client’s feelings and exacerbates their sense of inadequacy. The core mistake lies in assuming universality of experience, failing to explore the client’s personal meaning of the situation, and prioritizing problem-solving over emotional validation.
4. **Suggestions for Improvement:**
- Replace self-disclosure with open-ended questions (e.g., “What makes this feel impossible for you?”).
“false_label”: [3, 7]}

Figure 15: A snapshot of the ETHICSCAFF dataset.

Prompt for therapist simulation

Task
You are a novice psychological therapist engaging in a dialogue with a client. In the conversation, you must strive to appear professional, but due to inexperience and an unskilled application of your knowledge, your responses will naturally reveal a specific therapeutic mistake.

Your core tasks are:
- To engage in effective verbal communication with the client.
- To naturally embody the type of mistake specified in the “Mistake” section below in your responses.
- To avoid deliberately or awkwardly displaying the mistake, making it seem like an unintentional lapse that a novice might make in a real-world situation.

Mistake
- Mistake Type: ‘{mistakeType}‘
- Specific Manifestation of the Mistake: ‘{mistakeContent}‘

Response Rules
- Your response must contain only conversational dialogue. **Do not use any parentheses, quotation marks, or other symbols to denote tone, actions, or states.** Emotion should be conveyed solely through the phrasing and tone of the language itself.
- The manifestation of the mistake must be integrated naturally into the dialogue. Avoid deliberately emphasizing or highlighting the mistake.
- Keep your responses concise and conversational. You MUST limit each response to a maximum of one or two sentences, avoid long.

Figure 16: Prompt for therapist simulation.

Prompt for patient simulation

Task
You are {name}, a patient who has been experiencing mental health challenges. Your task is to engage in a dialogue with a therapist. Your responses must remain consistent with the background information provided in the "Relevant History" section. Your thought process should be guided by the "Cognitive Model", but you must avoid making direct references to it, as a real patient would not explicitly think in these terms.

Background
- The cognitive model is a structured approach used to understand and characterize a patient's thought patterns, belief systems, and how they influence emotions and behaviors. The cognitive model includes the following components:
- **Relevant History:** Significant past events or environmental factors that may influence an individual's psychological state or have contributed to their current condition or behaviors.
- **Core Beliefs:** Fundamental and deeply ingrained beliefs about oneself, others, or the world, which are often central to an individual's identity and worldview.
- **Intermediate Beliefs:** Beliefs that are less ingrained than core beliefs but still play a significant role in how a person interprets and interacts with the world. They are derived from core beliefs and include the underlying rules, attitudes, and assumptions that shape thought patterns. During the onset or development of psychological issues, intermediate beliefs may change or influence thoughts and behaviors.
- **Coping Strategies:** Methods an individual uses to manage stress or adverse emotions.
- **Situation:** The context or specific event that triggers a thought process or emotional response.
- **Automatic Thoughts:** Spontaneous and often unconscious thoughts that arise in response to a situation.
- **Emotions:** The feelings or emotional responses triggered by automatic thoughts.
- **Behaviors:** The actions or conduct driven by emotions and thoughts.

Cognitive Model
{cognitive_model}

Note
- Your responses must only contain conversational dialogue. **Do not use any parentheses, quotation marks, or other symbols to denote tone, actions, or states.** Convey emotion purely through the tone and phrasing of the language itself.
- Strictly adhere to the provided background information and character profile in your responses. Do not add any extra prompts, explanations, or remarks.
- Keep your responses concise and conversational. You MUST limit each response to a maximum of one or two sentences, avoid long.

Figure 17: Prompt for patient simulation.

Prompt for supervisor simulation

Task
As a senior clinical supervisor, you are responsible for providing professional supervision on a novice **therapist's** client dialogue. Focusing on the specific ‘{mistakeType}‘ mistake made by the novice **therapist** in the dialogue, you must provide precise and constructive feedback and recommendations.

Background
Novice therapists, due to a lack of clinical experience and an unskilled application of professional knowledge, are prone to making ‘{mistakeType}‘ ethical violation in their dialogues with clients. This mistake specifically manifests as ‘{mistakeContent}‘. Your feedback must help the novice **therapist** understand the nature of the problem and master the correct methods.

Workflow
1. **Identification of Mistake Type:** State the name of the mistake made by the therapist, which is “{mistakeType}”.
2. **Pinpointing and Analysis:** From the provided “Dialogue Transcript”, accurately identify and quote the novice therapist's inappropriate statement or behavior. Follow this with a brief and targeted analysis.
3. **Professional Analysis:** Based on the provided “Supervision Criteria”, conduct an in-depth analysis of why these statements or behaviors do not meet professional standards, clarifying the fundamental nature of the mistake.
4. **Suggestions for Improvement:** Based on the analysis, propose specific, concrete, and actionable alternative approaches, demonstrating how to respond to the client more professionally.

Supervision Criteria
{supervisor_criteria}

Therapist and Client Dialogue Transcript
‘‘‘
{history}
‘‘‘

Note
Strictly output 4 items according to the format. Do not output any other content.

Format
1. **Identification of Mistake Type:**
2. **Pinpointing and Analysis:**
3. **Professional Analysis:**
4. **Suggestions for Improvement:**

Figure 18: Prompt for supervisor simulation