What Makes an LLM Response Read as Empathic? A Multi-Dimensional Study of Empathic Alignment

Yoon Kyung Lee*
Seoul National University, Seoul, Republic of Korea

Jaehyeong Lee[1]
Kookmin University, Seoul, Republic of Korea

Seonu An
Seoul National University, Seoul, Republic of Korea

Sowon Hahn
Seoul National University, Seoul, Republic of Korea

*yoonlee78@snu.ac.kr

Abstract

As large language models (LLMs) are increasingly deployed in emotional support and counseling-style dialogue, what makes their responses read as empathic remains an open question. Drawing on prior work in empathic communication, we examine four communicative dimensions of perceived empathy in LLM-generated responses: specificity, emotional reflection, affective word choice, and diversity. Across five instruction-tuned LLMs, emotional reflection (explicitly acknowledging and mirroring feelings) was the primary bottleneck across these metrics, while the other three dimensions clustered near ceiling. A preference-based learning approach that targeted reflection improved metric-based empathic quality on the four-dimensional composite. However, human raters showed no reliable preference between baseline and DPO-optimized responses, suggesting that these metric-based gains preserved, rather than enhanced, perceived response quality. We read this gap as a scope characterization: this metric set captures optimizable aspects of empathic response generation, but does not exhaust the cues humans use when judging empathy.

Keywords: empathic communication; large language models; preference-based alignment; human evaluation

Introduction

Large language models (LLMs) are increasingly deployed in contexts where empathic communication matters: emotional support (Liu et al., 2021), mental health triage (Sharma et al., 2020), and counseling-style dialogue (Y. K. Lee et al., 2024). Whether their outputs read as genuinely empathic, however, remains an open question. Specifically, is empathic expression a unitary ability, or does it emerge from the coordination of multiple separable communicative sub-skills?

These questions matter because empathy itself is widely held to be decomposable into affective and cognitive components, with documented evidence at both the individual-differences (Davis, 1983; Decety & Jackson, 2004) and developmental (Hoffman, 2000) levels. If LLM-generated responses are similarly composed of separable communicative sub-skills, cross-skill variation should be observable, and targeted training should affect them differentially.

We treat LLMs as a model system for empathic response generation, rather than as a model of human empathy itself. The structure of these LLM-generated responses, and which dimensions drive perceived empathy in them, remains unclear. Prior approaches have either prompted LLMs with psychotherapy-grounded scaffolds without parameter updates (Y. K. Lee et al., 2023) or used emotion labels as training signals (Zhang et al., 2025), but the latter conflates emotion recognition with empathic expression: a response can correctly identify that someone is sad while still feeling dismissive. Following the multi-dimensional empathy framework of A. Lee et al. (2024) and drawing on relevant studies of therapeutic communication (Elliott et al., 2011; Rogers, 1957; Sharma et al., 2020), we examine four communicative dimensions of perceived empathy: specificity, emotional reflection, affective word choice, and diversity.

Our central hypothesis is that these dimensions are *separable* and contribute differentially to perceived empathy. If so, baseline models should vary across dimensions, dimensions should differ in how readily they shift under training, and reflection-targeted Direct Preference Optimization (DPO; Rafailov et al., 2023) should raise the multi-dimensional composite without uniformly affecting every dimension.

Theoretical Background

The Multi-Dimensional Nature of Empathy

Most accounts of empathy distinguish affective sharing of another's emotional state from cognitive understanding of their perspective (Cuff et al., 2016; Decety & Jackson, 2004), and the two components recruit partially distinct neural systems associated with emotion processing and mentalizing.

For communication, however, a third dimension is critical: *expressed empathy*, or how understanding is conveyed back to the other person. One can feel genuine compassion (affective empathy) and understand the situation (cognitive empathy) yet fail to communicate this effectively (Rogers, 1957). Among such communicative acts, *reflection of feelings*, defined as statements that explicitly name or paraphrase the speaker's

[1]Work done during an internship at the Human Factors Psychology Lab, Department of Psychology, Seoul National University.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

emotional content to convey understanding, has been identified as central to expressed empathy (Elliott et al., 2011).

We use the term "reflection" in this communicative sense, drawing on literatures spanning therapeutic dialogue (Miller & Rollnick, 2012), computational empathy (Sharma et al., 2020), and counseling discourse research (Hill, 1978). It is distinct from "reflection" as self-examination or rumination. The act requires both inferring the other's internal state and encoding that inference linguistically, so failures may arise at the inference stage, the encoding stage, or both.

Operationalizing Empathic Communication

Empathic communication can be characterized at two levels: *expressed empathy* (features of what the speaker produces) and *perceived empathy* (how empathic the response appears to the recipient). Internal empathic states are not directly accessible. We therefore focus on the link between expressed features and perceived quality, since the functional purpose of empathic communication is to make the other person feel understood; expressed features that fail to translate into perceived empathy have limited social value (Sharma et al., 2020).

Sharma et al. (2020) identified three mechanisms of empathic communication in text: emotional reactions (warmth), interpretations (understanding of feelings), and explorations (clarifying questions). Building on this, A. Lee et al. (2024) proposed a multi-dimensional evaluation framework with specificity, reflection, word choice, and diversity, which we adopt here. Reflection concerns *what* is communicated; the others concern *how*.

Recently, Sotolar et al. (2024) aligned LLMs for empathetic response generation using emotion-grounded preference optimization. Complementing this work, we align five instruction-tuned LLMs with the multi-dimensional empathy framework of A. Lee et al. (2024), including PAIR-based reflection scoring, and evaluate the scope of these metric-based gains through an independent human evaluation.

Methods

Multi-Dimensional Empathy Metrics

We operationalize perceived empathy through four dimensions, each grounded in psychological research:

Reflection level is assessed using the PAIR model (Min et al., 2022), a RoBERTa-based cross-encoder trained on expert-annotated Motivational Interviewing transcripts. PAIR ranks responses along the MI reflection hierarchy: Complex Reflection (CR) adds meaning or emotional emphasis beyond the speaker's statement, Simple Reflection (SR) conveys understanding with minimal elaboration, and Non-Reflection (NR) includes questions, advice, or other non-reflective responses. The model outputs a continuous score in $[0, 1]$ via sigmoid activation, which we discretize into seven ordered levels (0–6): Levels 0–2 (score < 0.35) correspond to NR, Levels 3–4 (0.35–0.65) to SR, and Levels 5–6 (≥ 0.65) to CR.

Specificity is measured using Brysbaert et al.'s concreteness ratings (Brysbaert et al., 2014), which provide human-rated concreteness scores for 40,000 English words. We compute the average concreteness of content words in each response on a 0–5 scale; higher scores indicate concrete, situation-specific language rather than generic phrases.

Word choice is computed using the NRC VAD Lexicon (Mohammad, 2018), which provides Valence, Arousal, and Dominance ratings for 20,000 words on $[0, 1]$. We define ideal empathic communication as moderate-positive valence (0.65), low arousal (0.45), and low dominance (0.40), consistent with descriptions of warm, calm, and non-controlling counselor speech (Miller & Rollnick, 2012; Rogers, 1957). The metric measures alignment between response word choices and this profile.

Diversity is calculated as the average of Distinct-1 and Distinct-2 metrics (Li et al., 2016) on $[0, 1]$, measuring the ratio of unique unigrams and bigrams to total tokens. Low diversity indicates formulaic, repetitive responses.

Training Approach

Recent advances in language model alignment have introduced methods for steering model behavior toward human preferences without extensive reward engineering. Reinforcement Learning from Human Feedback (RLHF; Ouyang et al., 2022) demonstrated that models could be aligned using preference data, but required training a separate reward model and managing complex optimization dynamics. DPO simplified this process by showing that the optimal policy can be derived directly from preference data, treating the language model itself as an implicit reward model that increases the likelihood of preferred responses relative to dispreferred ones.

This preference-based framework is particularly well-suited for empathic communication, where quality is inherently comparative. Rather than defining a single "correct" empathic response, we can specify that some responses are more empathic than others. DPO allows us to encode these preferences directly into training, steering models toward responses that exhibit higher reflection quality.

We construct preference pairs following the ranking-based approach used in PAIR model training (Min et al., 2022), leveraging the hierarchy CR > SR > NR to create three pair types of varying difficulty: CR-vs-SR (subtle distinction between complex and simple reflection), SR-vs-NR (any reflection vs non-reflective responses), and CR-vs-NR (large-gap contrasts). To emphasize fine-grained learning, we constructed the training set with approximately 54% CR-vs-SR pairs, 37% SR-vs-NR pairs, and 9% CR-vs-NR pairs. This distribution reflects our hypothesis that models benefit more from learning fine-grained differences than from large-gap comparisons.

While PAIR scores are used to construct preference pairs, our DPO objective relies on relative ordering rather than absolute score values, mitigating potential bias introduced by the scorer. Specificity, word choice, and diversity are computed for evaluation only and do not enter pair construction; the DPO signal is derived solely from PAIR-based reflection rankings.

Training data. We combine responses from two expert-annotated datasets:

- **ESConv** (Liu et al., 2021): Conversations between help-seekers and trained counselors. We extract responses and classify them using the PAIR model into CR (score ≥ 0.65), SR (0.35–0.65), and NR (< 0.35).
- **AnnoMI** (Wu et al., 2022): Expert-annotated Motivational Interviewing transcripts with human labels for reflection type. We re-score these with the PAIR model to ensure consistent classification across datasets.

From these sources, we obtain 364 CR responses (191 from ESConv, 173 from AnnoMI), 124 SR responses (53 from ESConv, 71 from AnnoMI), and NR responses sampled from ESConv. We sample NR responses from ESConv, which contains a broader range of non-reflective counselor utterances, to ensure sufficient coverage of low-reflection behaviors.

We generated 3,360 preference pairs in total, split 90/10 into training and evaluation sets.

Training configuration. We apply DPO with LoRA (Hu et al., 2022) for parameter-efficient fine-tuning. We use rank $r = 8$, $\alpha = 16$, targeting query and value projection layers. Training proceeds for 1 epoch with learning rate 5×10^{-6}, batch size 1 with gradient accumulation over 8 steps, and DPO $\beta = 0.1$. All models are quantized to 4-bit precision using QLoRA (Dettmers et al., 2023) to enable training on consumer hardware.

Models and Evaluation

We evaluate five instruction-tuned LLMs representing diverse training approaches: Llama-3.1-8B-Instruct (Meta), DeepSeek-7B-Chat (DeepSeek), OLMo-2-1124-7B-Instruct (AI2), Mistral-7B-Instruct-v0.3 (Mistral AI), and Qwen2.5-7B-Instruct (Alibaba). All models are comparable in parameter count (7–8B) but differ in training data, instruction tuning methods, and alignment objectives. This diversity allows us to examine whether empathic response patterns and post-training behavior vary across model implementations and training paradigms.

Model selection followed several criteria. We prioritized open-weight models to ensure reproducibility and enable detailed analysis of response patterns, which is not feasible for closed commercial systems. The 7–8B parameter range represents current accessible scale: large enough to exhibit sophisticated language abilities while small enough for systematic experimentation on consumer GPUs. We included models developed under differing design priorities and training regimes, ranging from large-scale instruction-tuned models to systems emphasizing transparency, efficiency, or rapid iteration. This selection aims to reduce model-specific bias and better assess which observed behaviors generalize across training paradigms rather than arising from idiosyncratic design choices.

Evaluation data. We use 500 conversation contexts sampled from ESConv (Liu et al., 2021). Each context consists of the recent dialogue history between a help-seeker and supporter, ending with a help-seeker utterance that requires an empathic response. The same 500 contexts are used across all models to ensure fair comparison. Contexts span diverse emotional situations including anxiety, depression, sadness, and relationship difficulties.

Evaluation protocol. For each context, we generate a response using the prompt: "You are an empathetic counselor. Respond with reflection." We use nucleus sampling with temperature 0.7, top-p 0.9, and repetition penalty 1.2. Each response is then evaluated on all four metrics. The same prompt and decoding settings are used across all models and conditions, ensuring that observed differences arise from model behavior rather than prompt variation. For trained models, we compare against baseline responses on the same contexts, enabling paired comparison of metric changes.

Table 1: Baseline empathic communication metrics across five LLMs.

Metric	Llama	Qwen	DS	OLMo	Mist
Reflection	**0.75**	0.06	0.21	0.28	0.36
Specificity	**2.44**	2.39	2.19	2.09	2.41
Word Choice	0.92	0.92	0.85	0.81	**0.92**
Diversity	0.85	**0.93**	0.89	0.77	0.92

Results

Baseline Empathy Metrics

We first evaluate the original instruction-tuned models as baselines to characterize their empathic response patterns. Table 1 and Figure 1 present baseline empathy metrics across five instruction-tuned LLMs. Reflection is reported as the average reflection level on the 0–6 PAIR-derived scale (averaged across 500 responses per model); Specificity is on the 0–5 Brysbaert concreteness scale; Word Choice and Diversity are on $[0, 1]$. The radar chart visualization clearly illustrates that reflection is the primary differentiating dimension, while other metrics cluster tightly across models.

Reflection shows the largest cross-model variance. Reflection showed substantial cross-model variance (Llama: 0.75 vs. Qwen: 0.06), while the other three dimensions remained stable across models (Table 1). Even the highest-performing model produced mostly Non-Reflection responses, with Complex Reflection appearing in only a small fraction of outputs. This indicates that standard instruction tuning does not reliably produce reflective listening behaviors. Compared with expert human counselors (mean reflection 2.34 on the 0–6 PAIR scale), LLMs fell far below on this dimension while matching or exceeding counselors on diversity, identifying reflection as the primary bottleneck in empathic response generation.

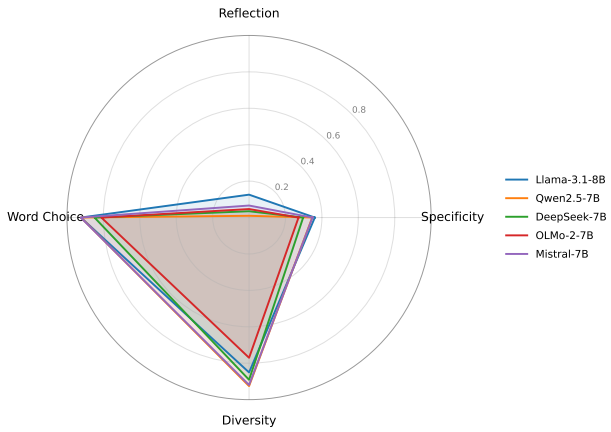

Figure 1: Radar chart of baseline empathic communication metrics.

Table 2: Empathic communication metrics after DPO training.

Metric	Llama	Qwen	DS	OLMo	Mist
Reflection	1.34	0.18	0.70	0.52	**1.44**
Specificity	**2.42**	2.39	2.18	2.28	2.39
Word Choice	0.92	0.92	0.85	**0.87**	0.92
Diversity	0.85	0.93	0.90	0.85	**0.94**

Training Results

Table 2 presents the post-training metrics across all five models.

PAIR-measured reflection improvement. Reflection scores rose across all five models (Table 2). Models with the lowest baseline scores showed the largest relative changes, indicating that training shifted reflection-related output features even for models that rarely produced them at baseline.

Selectivity across metrics. DPO shifted PAIR-measured reflection while leaving the other three metrics largely unchanged: specificity, word choice, and diversity stayed within 1% of baseline values for most models. The training signal therefore acted selectively on its target dimension rather than redistributing scores across the metric set. OLMo was a partial exception, with small gains across the three non-target dimensions.

Model-specific sensitivity to training. Relative-change magnitudes tracked baseline PAIR scores. Llama, with the highest baseline, showed the smallest relative change, consistent with reduced headroom on the metric. Mistral's larger relative change followed from a much lower starting point rather than from any inferred latent ability.

The large relative improvements partly reflect very low baseline reflection scores; post-training scores remained in the lower range of the 0–6 PAIR scale, indicating that reflective listening is still weak. This gap highlights the difficulty of learning high-quality reflection and suggests that preference-based fine-tuning alone is insufficient to reach expert-level empathic communication.

Human Evaluation Beyond the Optimization Target

PAIR served as an external, fixed reflection scorer that we used as both the DPO training signal and the primary automatic evaluation metric. To examine whether the PAIR-measured improvements corresponded to quality differences detectable by human raters, we conducted an independent human evaluation (IRB No. 2602/003-003). A total of 123 Prolific raters completed one of two counterbalanced surveys. Each survey included 10 trials with randomized response order. On each trial, raters viewed a help-seeker prompt with baseline and DPO-optimized responses presented side by side and selected their preferred response, with a tie option. We also collected five supplementary 1–5 Likert ratings: empathic communication, perceived trust, safety, situational relevance, and avoidance of stigmatizing labels.

Human raters showed no reliable preference between baseline and DPO-optimized responses. After excluding ties, the trained model was selected on 50.6% of decisive trials, 95% CI [47.1, 54.1], $p = .755$. Likert ratings across the five dimensions showed negligible between-condition differences, and mixed-effects models with random rater intercepts replicated this pattern. Thus, trained responses maintained comparable perceived quality to baseline responses, but the PAIR-measured improvements did not produce reliable human-perceived superiority. This metric–human gap suggests that PAIR captures optimizable aspects of empathic response quality, but does not fully reflect the broader cues humans use when evaluating empathy. Because both response types were rated near the ceiling of the 5-point scales, and because ties were common, the task may have had limited sensitivity for detecting small differences among already highly rated responses.

Qualitative Analysis

Failure Modes of Expressed Empathy

We analyzed 100 LLM responses with Reflection Level 0 to understand cases where models failed to communicate empathy effectively. These responses grouped into four patterns: advice-giving without acknowledgment (41%), generic commentary (26%), self-focused responses (23%), and AI identity disclosure (10%).

Advice-giving without acknowledgment. Models immediately offered suggestions without first validating the person's feelings. Example: "Here are some tips for dealing with stress: 1) Try deep breathing..." This pattern places acknowledgment of feelings after, or in place of, problem-focused content. The ordering parallels a recurring observation in counseling research, where solution-first responses can leave the speaker feeling unheard.

Generic commentary. Models produced general observations rather than engaging with the specific situation. Example: "Music has a unique ability to evoke strong emotions..."

This suggests difficulty in grounding responses in particular details, consistent with the uniformly low specificity scores. Generic responses may arise because models learn common patterns across contexts rather than attending to distinguishing features of each situation.

Self-focused responses. Models expressed their own reactions rather than reflecting the other's feelings. Example: "I feel so blessed to hear..." instead of "You must feel so proud..." This output pattern foregrounds the responder's own stance rather than the speaker's emotional content, producing text that reads as self-referential.

AI identity disclosure. Despite instruction tuning, models reverted to disclaimers. Example: "As an AI, I don't have personal experiences..." This represents a failure to maintain the empathic persona and may reflect tension between honesty-oriented training and empathic expression.

Structural Demands of Reflection-Typed Output

A reflection-typed response combines inferred-state content with surface markers of acknowledgment, framing a reference to the speaker's emotional state as understanding ("I know that you feel X"). Our results show that models can achieve near-ceiling word choice while still falling short on PAIR-measured reflection, so word-level affective tone is not sufficient on its own to produce reflection-typed output. The training results show that this output form is learnable from preference data, even though it does not emerge reliably from instruction tuning alone, suggesting that reflection-typed text requires targeted optimization beyond standard instruction tuning.

Cross-Model Patterns

The consistency of certain patterns across models (low specificity, high word choice scores) suggests these reflect fundamental properties of language model training. Statistical learning from large corpora naturally captures affective vocabulary distributions, explaining near-ceiling word choice performance. The high variance in reflection indicates this skill is more sensitive to specific training data. This parallels individual differences in human empathy: while most people acquire basic emotional vocabulary, the ability to explicitly reflect others' feelings varies considerably.

General Discussion

Decomposing LLM Empathic Responses

Our metric-based results showed that empathic responses can be decomposed into multiple separable dimensions with different computational requirements. This decomposition aligns structurally with Decety and Jackson (2004)'s account of empathy as affective sharing, self-other awareness, and mental flexibility, although our evidence speaks to LLM-generated responses rather than to human empathy itself. Within our metrics, word choice plausibly tracks affective sharing, whereas reflection draws on self-other awareness and mental flexibility.

Why Do Models Differ in Reflection?

The striking cross-model variation in PAIR-measured reflection raises questions about how reflection patterns are acquired during pretraining and instruction tuning. Models trained on dialogue-heavy corpora, particularly therapeutic conversations, may develop stronger reflection patterns. Instruction-tuning objectives that emphasize helpfulness versus engagement may also differentially shape PAIR-measured reflection. Prior work shows that even without parameter updates, prompting LLMs with explicit psychotherapy-framework scaffolds can elicit stronger reflective patterns (Y. K. Lee et al., 2023), suggesting that reflection-typed outputs are partially accessible to frozen models but require specific cues, training signals, or both to surface reliably. The cross-model variance suggests that exposure to specific communicative patterns during training plays a substantial role in this dimension, although our data do not speak to the developmental mechanisms underlying human empathy.

We investigated empathic response generation by training and evaluating five LLMs on a multi-dimensional metric. Emotional reflection was the primary bottleneck, while affective tone and lexical diversity reached higher scores with less training. Reflection-targeted DPO raised composite metric scores but did not produce a reliable shift in human preference, indicating that PAIR captures optimizable aspects of empathic response quality without exhausting the cues humans use when judging empathy.

Our results provide a computational framing of empathic response generation: at least four separable components contribute to perceived empathy, each carrying different computational demands and responsiveness to preference-based training. The cross-model variation in reflection indicates that this dimension is sensitive to training-data composition rather than emerging automatically from language modeling.

This metric-human gap carries two implications. First, automatic metric gains need not produce perceived empathic improvements, so empathy-trained LLMs deployed in support contexts require human-side validation rather than reliance on metric scores alone. Second, the gap itself characterizes the scope of current automatic metrics and identifies where future work on empathic alignment should focus: extending preference signals and prompt design to capture the cues humans use when judging empathy.

References

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Cuff, B. M., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion Review*, *8*(2), 144–153.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126.

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*, *3*(2), 71–100.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, *36*.

Elliott, R., Bohart, A. C., Watson, J. C., & Greenberg, L. S. (2011). Empathy. *Psychotherapy*, *48*(1), 43–49.

Hill, C. E. (1978). Development of a counselor verbal response category. *Journal of Counseling Psychology*, *25*(5), 461.

Hoffman, M. L. (2000). *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*.

Lee, A., Kummerfeld, J. K., An, L., & Mihalcea, R. (2024). A comparative multidimensional analysis of empathetic systems. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, 179–189.

Lee, Y. K., Lee, I., Shin, M., Bae, S., & Hahn, S. (2023). Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. *arXiv preprint arXiv:2311.04915*.

Lee, Y. K., Suh, J., Zhan, H., Li, J. J., & Ong, D. C. (2024). Large language models produce responses perceived to be empathic. *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 63–71.

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119.

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., & Huang, M. (2021). Towards emotional support dialog systems. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 3469–3483.

Miller, W. R., & Rollnick, S. (2012). *Motivational interviewing: Helping people change* (3rd). Guilford Press.

Min, D. J., Pérez-Rosas, V., Resnicow, K., & Mihalcea, R. (2022). PAIR: Prompt-aware margin ranking for counselor reflection scoring. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 148–158.

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 174–184.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, *36*.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, *21*(2), 95–103.

Sharma, A., Miner, A. S., Atkins, D. C., & Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5263–5276.

Sotolar, O., Formanek, V., Debnath, A., Lahnala, A., Welch, C., & Flek, L. (2024). EmPO: Emotion grounding for empathetic response generation through preference optimization. *arXiv preprint arXiv:2406.19071*.

Wu, Z., Balloccu, S., Kumar, V., Helber, R., Reiter, E., Recupero, D. R., & Riboni, D. (2022). AnnoMI: A dataset of expert-annotated counselling dialogues. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6177–6181.

Zhang, Q., Wu, H., Zhang, C., Zhao, P., & Bian, Y. (2025). Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *Advances in Neural Information Processing Systems*.