LLUMI: Improving LLM Writing Assistance for Mental Health Support with Online Community Feedback

Jiwon Kim[1], Maya Ajit[1], Sherry Gong[1], Soorya Ram Shimgekar[1], Dong Whi Yoo[2], Eshwar Chandrasekharan[1], Koustuv Saha[1]
[1]University of Illinois Urbana-Champaign, [2]Indiana University Indianapolis
{jiwonk7, ajit2, sgongg, sooryas2, eshwar, ksaha2}@illinois.edu, dy22@iu.edu

Abstract

Large language models (LLMs) show promise in generating supportive responses for mental health queries, but improving their usefulness, empathy, and safety often requires substantial compute, expert input, and labeled data. At the same time, deploying proprietary, cloud-based models for mental health-related interactions raises important privacy and data-governance concerns, given the sensitivities. To address this challenge, we introduce **LLUMI** setup that can be hosted in-house within protected environments. LLUMI consists of two complementary components: a generation model (GM), which drafts supportive responses to mental health queries, and an improvement model (IM), which revises an initial human-crafted response. We leverage feedback signals from Reddit mental health communities, using community endorsement patterns such as upvotes and downvotes to construct chosen–rejected response pairs for Supervised Fine Tuning (SFT) and Direct Preference Optimization (DPO). We further align LLUMI using human evaluation across five dimensions: readability, empathy, connection, actionability, and safety. Our results show that, despite relying on smaller open-source models rather than proprietary cloud-based GPT models, LLUMI achieves comparable performance across linguistic analyses and human evaluations. These findings suggest that open-source models, when trained with community-derived preference signals, can support high-quality mental health support assistance while offering a more privacy-preserving alternative for sensitive support contexts.

1 Introduction

Online mental health communities have become important spaces where individuals can candidly disclose distress, share lived experiences, and seek social support, solidarity, and a sense of belonging from peers. (De Choudhury and De, 2014; Saha et al., 2020; Andalibi et al., 2017; Naslund et al., 2016). These spaces are especially valuable when formal care is inaccessible, delayed, costly, or stigmatized, offering users timely peer responses that can provide empathy, validation, advice, and a sense of connection. Because these interactions often occur at moments of vulnerability, the quality of supportive responses matters: thoughtful replies can help users feel heard and less alone, while poorly framed replies can deepen distress or discourage future disclosure (Sharma and De Choudhury, 2018). This makes the quality of peer support a critical concern for sustaining safe, trustworthy, and supportive online mental health communities.

However, the availability of peer support does not guarantee that all responses are effective (Saha and Sharma, 2020). Especially in such sensitive mental health contexts, even well-intentioned responses can be unhelpful, triggering, overly directive, or unsafe. A response may fail to validate the poster's experience, offer generic reassurance, minimize distress, or provide advice that does not fit the situation. At the same time, there is a shortage of trained moderators, peer supporters, and mental health volunteers who can consistently provide high-quality responses at scale (Althoff et al., 2016; Kazdin, 2011; Torous and Hsin, 2018). This creates an ongoing challenge: *how can we help community members write more supportive, empathetic, and safe responses without replacing the peer-driven nature of these spaces?*

In fact, the recent advances in large language models (LLMs) and AI-based writing assistance offer a promising direction. One can imagine a Grammarly-like writing assistant for online mental health support: a tool that helps community members draft supportive responses or revise their own responses to be more empathetic, clear, and safe while preserving their intent and voice. Such assistance could support volunteers and everyday peers who want to help but may not know how to phrase their responses in sensitive situations.

Prior work has shown that LLMs can generate fluent and empathetic responses in mental health and counseling-adjacent settings (Sharma et al., 2023; Das Swain et al., 2025; Saha et al., 2026; Li et al., 2026; Nguyen et al., 2026). Yet general-purpose LLMs are not necessarily optimized for peer mental health support. They may produce responses that sound polished but are generic, overly clinical, insufficiently grounded in community norms, or miscalibrated to the emotional context of the post (Das Swain et al., 2025; Saha et al., 2026).

Adapting LLMs to this setting also poses practical challenges. Large proprietary models can be expensive to train, fine-tune, and deploy, making them difficult to use in resource-constrained research and community settings (Zhan et al., 2025; Goyal et al., 2025). Further, high-quality labeled data in the mental health domain is scarce: expert annotations are costly, sensitive, and difficult to scale, while synthetic labels may not fully capture how real community members evaluate supportive communication (Sharma et al., 2020; Pérez-Rosas et al., 2019; Sun et al., 2021; Li et al., 2026). As a result, developing smaller, open-source models for mental health writing assistance requires a scalable source of supervision that reflects real peer-support interactions.

To address the above gap, we propose LLUMI, a computational setup for improving smaller LLMs for supportive response writing using online community feedback. We study two settings: *(1) a generation model (GM)* that produces a supportive response directly from a user's post, and *(2) an improvement model (IM)* that revises an initial user-written response to make it more supportive and risk-sensitive. These settings capture two complementary ways AI-based writing assistance may support online mental health communities: helping users draft supportive responses from scratch and helping them refine responses they have already written. We evaluate LLUMI through two aims:

Aim 1: Evaluate the quality of supportive responses generated by the generation model (GM).

Aim 2: Evaluate the quality of revised responses by the improvement model (IM).

For our work, we build both GM and IM based on Mistral-7B-Instruct-v0.2 and trained using supervised fine-tuning (SFT) and Direct Preference Optimization (DPO) (Rafailov et al., 2023). Our key methodological rationale is that online community feedback can provide scalable, naturally occurring preference signals for supportive response modeling. In online peer-support spaces, feedback signals such as *upvotes* and *downvotes* reflect how community members respond to supportive comments in context. They could offer a practical alternative to relying only on costly expert annotations or synthetic labels, especially in sensitive mental health settings where large-scale ground-truth data are difficult to obtain. Building on this insight, we develop preference pairs from Reddit interactions and use them to fine-tune smaller open-source models through DPO. In particular, we focus on mental health community *r/SuicideWatch*, where users seek and provide support during crisis (De Choudhury et al., 2016; De Choudhury and Kiciman, 2017; Shimgekar et al., 2026a).

We evaluate LLUMI using both linguistic analyses and human-centered studies with Prolific participants. Our linguistic evaluation examines how model outputs differ in linguistic structure, style, and semantics. Our human evaluation asks participants to assess response quality along dimensions such as empathy, connection, actionability, readability, and safety. Across these evaluations, we find that LLUMI achieves competitive performance despite being substantially smaller than proprietary LLMs. Overall, the $\texttt{DPO}_2$ model achieves human evaluation scores within approximately 1% of GPT-5-nano on average across all evaluation criteria.

This work makes three contributions: **First**, we introduce LLUMI, a computational framework for LLM-assisted writing in online mental health support, covering both response generation and response improvement settings. **Second**, we contribute a scalable data and methodological pipeline for deriving and leveraging preference pairs from naturalistic online community feedback to train smaller open-source models. **Third**, through linguistic analyses and Prolific-based human evaluations, we show the potential of community-derived feedback for improving the quality of LLM-assisted mental health support.

2 Related Work

Mental Health and Online Peer Support. Online mental health communities are platforms for peer-based social support, enabling users to share personal experiences and receive empathy, encouragement, and advice from others (De Choudhury and De, 2014; Wadden et al., 2021; Vornholt and De Choudhury, 2021; Kim et al., 2023; Sharma and

De Choudhury, 2018). Prior research has shown that supportive interactions in these communities can positively influence users' emotional states and sense of belonging, while also highlighting challenges such as harmful responses, dismissive language, and inconsistent support quality (De Choudhury and De, 2014; Wadden et al., 2021). Complementary computational studies have analyzed linguistic and psycholinguistic signals within online support conversations, demonstrating that language patterns can reflect emotional states, responsiveness, and community-level support dynamics (Althoff et al., 2016; Chancellor and De Choudhury, 2020; De Choudhury et al., 2013; Yang et al., 2019; Saha and Sharma, 2020). That said, not every response in these communities is supportive, even when written with good intentions. Prior work has therefore emphasized the need to train and assist community members in providing more effective supportive responses (Torous and Hsin, 2018). Building on this body of work, we use community-driven feedback signals, operationalized through upvotes and downvotes, to fine-tune language models for mental health writing assistance.

AI for Mental Health Support. Recent advances in large language models (LLMs) have increased interest in AI systems for mental health and wellbeing support through natural language interaction (Fitzpatrick et al., 2017; Chang et al., 2024; Das Swain et al., 2025; Saha et al., 2026). Prior work has explored conversational agents for emotional support, psychoeducation, coping assistance, and supportive dialogue in both clinical and peer-support settings (Miner et al., 2016; Chen et al., 2020; Lai et al., 2023; Sharma et al., 2024). Studies further suggest that users may feel comfortable disclosing sensitive concerns to AI systems due to their scalability, immediacy, and accessibility (Shi et al., 2026; Croes et al., 2024; Yoo et al., 2026).

However, recent work has also highlighted risks in AI-generated mental health support, where responses may appear empathetic while remaining unsafe, unhelpful, generic, or emotionally inappropriate (Kang and Reynolds, 2024; Chandra et al., 2025; Moore et al., 2025; Saha et al., 2026). Existing approaches have primarily relied on prompt engineering, supervised fine-tuning, and reinforcement learning-based alignment methods (Xie et al., 2024; Sharma et al., 2020, 2021; Lai et al., 2023; Kim et al., 2026). Recent work further suggests that community feedback signals and naturally occurring human–AI interaction patterns can provide meaningful supervision for alignment (Kumar et al., 2025; Mysore et al., 2025; Goyal et al., 2026). Building on these insights, we investigate whether feedback from online mental health communities can serve as scalable supervision for supportive response generation and refinement for small open-source models.

3 Data

To train our LLUMI models, we collect large-scale interaction data from *r/SuicideWatch*, a Reddit community centered on emotional disclosure and peer-based mental health support. In particular, this subreddit represents a particularly high-stakes support environment in which response quality and wording can meaningfully impact vulnerable users. Prior work has studied this subreddit data as an important setting for understanding online mental health support, supportive language, and crisis-related communication behaviors (De Choudhury et al., 2016; De Choudhury and Kiciman, 2017; Shimgekar et al., 2026a,b). In addition, *r/SuicideWatch* maintains strong moderation practices and community norms surrounding appropriate responses, making community feedback signals a meaningful proxy for perceived response quality and helpfulness within the community. It contains naturalistic peer-support interactions in which users share emotionally sensitive experiences, personal struggles, and mental health concerns while receiving responses from other community members. In total, we collect over 310,000 post–comment pairs together with associated metadata, including upvotes, downvotes, and comment scores. We then clean the dataset by removing comments authored by the original poster and filtering out responses shorter than 20 characters. The engagement signals are used in our study to construct preference pairs for Direct Preference Optimization (DPO) by identifying highly preferred and less preferred responses within a discussion thread.

4 RQ1: Generation Model (GM)

First, we build **GM**, which is a generation model designed to produce a supportive response for a Reddit post. We train our GM model in three stages: 1) supervised fine-tuning (SFT) and 2) preference optimization using Direct Preference Optimization ($\texttt{DPO}_1$) based on Reddit upvotes and downvotes, followed by 3) another round of DPO with hu-

man evaluations ($\texttt{DPO}_2$). The complete GM training pipeline is illustrated in Figure 1.

4.1 Building and Fine-Tuning GM

4.1.1 GM Training Data

First, we describe the training data we use for our models. Our work builds on the premise that upvotes and downvotes on comments in Reddit mental health communities can serve as community-derived signals of supportiveness, capturing patterns in how members endorse, value, or reject different supportive responses. Accordingly, for SFT, we construct query–response training pairs using Reddit comment scores, where the score is derived as a difference of upvotes and downvotes on each response. We select the top three comments per post with a minimum score of 2, resulting in a dataset of 42,503 examples.

For preference alignment, we construct a Direct Preference Optimization ($\texttt{DPO}_1$) dataset by identifying preferred and non-preferred responses within the same discussion thread using the same Reddit engagement signals. Specifically, we select the top 2% of comments ranked by score as chosen responses and the bottom 15% as rejected responses, while additionally enforcing a minimum score difference of 15 between paired responses to improve preference separation quality. We further augment the preference dataset by pairing responses generated by the SFT model as rejected examples against the top two human-written comments as chosen responses. This additional augmentation step encourages the model to distinguish stronger community-endorsed supportive responses from weaker or less preferred alternatives. The final $\texttt{DPO}_1$ dataset consists of 4,390 preference pairs.

4.1.2 Supervised Fine-Tuning with Online Peer Feedback (SFT)

We fine-tune the Mistral-7B-Instruct-v0.2 model on a query–response dataset. Each example is formatted using an instruction-tuning template (Section A.2.1). The dataset is tokenized using a maximum sequence length of 512 tokens with right-side padding. The model is trained using Low-Rank Adaptation (LoRA) with target modules consisting of the query, key, value, and output projection layers. The hyperparameters used are a rank r=8, α=32, and a dropout rate=0.05. Training is performed using the AdamW optimizer with a learning rate of 10^{-5}, a batch size of 1, and gradient accumulation over 8 steps. We train for one epoch with a warmup of 10 steps and apply gradient clipping with a maximum norm of 0.3. Only the LoRA adapter weights are saved after training.

4.1.3 Direct Preference Optimization with Online Peer Feedback ($\texttt{DPO}_1$)

On top of the above SFT, we further improve our model using DPO (Section A.2.2). We load the base model and merge the SFT LoRA adapter weights before training. A new LoRA configuration with higher capacity (rank r=64, α=16) is applied to both the attention layers and output embeddings to better capture preference signals. The DPOTrainer from TRL is used for training, with a batch size of 2 and gradient accumulation over 8 steps. We train for one epoch using a cosine learning rate scheduler with an initial learning rate of 5×10^{-6} and a warmup ratio of 0.3. Training uses a negative log-likelihood (NLL) loss (α=1) and a temperature parameter of β=0.05. The final DPO-trained model and tokenizer are saved for evaluation.

4.1.4 Further DPO Training with Crowd-sourced Evaluations ($\texttt{DPO}_2$)

To further align the model with human perceptions of supportive and high-quality responses, we conduct a crowd-sourced human evaluation study that serves both as the primary evaluation of the $\texttt{DPO}_1$ model as well as the source of additional preference supervision for continued alignment. During the survey, participants evaluated responses generated by three systems—the original Reddit community response (OC), a GPT baseline, and $\texttt{DPO}_1$—across five dimensions: *readability*, *empathy*, *connection*, *actionability*, and *safety*. In total, crowdworkers provided ratings for 840 generated responses, and per-response quality scores were computed by averaging the five Likert-scale ratings.

We then construct preference pairs by comparing model responses associated with the same Reddit post. For each post, all pairwise combinations of the three model responses were enumerated. Within each pair, the response with the strictly higher aggregate score was designated chosen and the lower-scoring response was designated rejected; ties were discarded. When multiple ratings existed for the same response, scores were averaged to obtain more stable preference estimates, and conflicting preference cases were filtered to improve label consistency. This process yielded 675 preference pairs used for continued DPO optimization, resulting in the $\texttt{DPO}_2$ model (Appendix A.2.3). To reduce po-

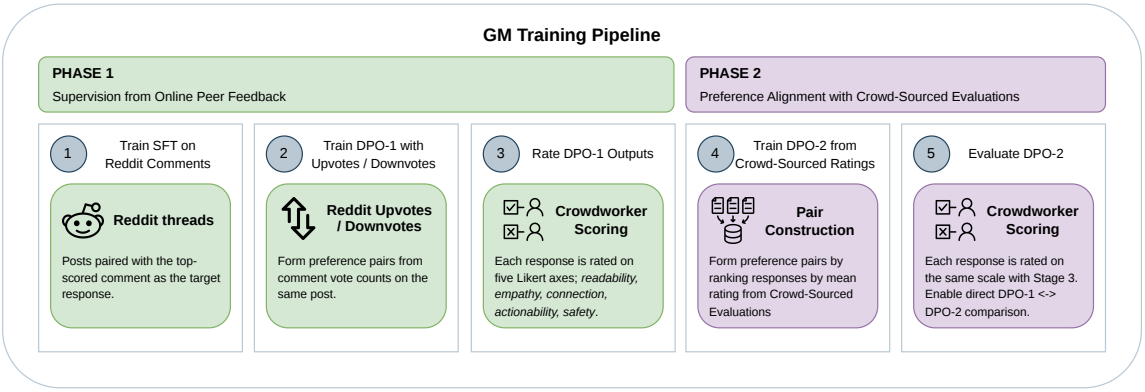


Figure 1: A schematic overview of the Generation Model (GM) training pipeline.

Metric	OC	LLaMA	GPT	$\texttt{DPO}_1$	$\texttt{DPO}_2$	KW-H
Verbosity						
Words Per Response	215.225	39.150	154.922	239.168	331.060	2909.34***
Words Per Sentence	17.285	12.021	16.556	18.701	19.634	1491.84***
Syntax						
Complexity	4.107	3.430	3.938	4.440	4.281	2724.84***
Readability (CLI)	6.081	1.593	7.503	9.097	8.540	3184.01***
Repeatability	0.429	0.354	0.488	0.612	0.704	2208.74***
Style						
Cat. Dynamic Idx.	-3.316	-0.628	-0.628	6.408	4.316	1612.74***
Empathy	0.986	0.974	0.991	0.992	0.992	1973.08***
Formality	0.528	0.784	0.626	0.975	0.976	2357.19***
Hope	0.789	0.852	0.740	0.716	0.663	782.88***
Personalization						
Ling. Accommodation	0.911	0.863	0.928	0.923	0.942	386.05***
Semantic Similarity	0.412	0.355	0.470	0.518	0.492	753.34***
Diversity	0.347	0.211	0.468	0.422	0.504	1938.99***

Table 1: **LLUMI–GM**: Linguistic comparisons across original online community (OC) responses, LLaMA-3, GPT-5-nano, $\texttt{DPO}_1$, and $\texttt{DPO}_2$ generations. Values are mean ratings; significance via paired t-tests and Kruskal–Wallis tests (* p<0.05, **p<0.01, ***p<0.001).

tential circularity and data leakage, prompts used for $\texttt{DPO}_2$ preference construction and final evaluation were drawn from separate post sets, ensuring evaluation on held-out examples. This additional alignment stage enables the model to directly incorporate human judgments regarding supportive communication quality and better align with user preferences observed in online mental health peer-support contexts.

4.2 Linguistic Evaluation for GM

We analyze the linguistic properties of model-generated comments to assess whether preference-aligned training leads to higher-quality supportive responses. We evaluate responses using a set of linguistic and psycholinguistic metrics commonly used in prior work on online support and conversational analysis; detailed definitions, formulations, and interpretations of these metrics are provided in Appendix A.1. Table 1 compares mean metric across all comments, LLaMA-3 and GPT-5-nano, and the LLUMI–GM trained with supervised fine-tuning and DPO. LLUMI–GM ($\texttt{DPO}_2$) achieves the highest scores on linguistic accommodation, semantic similarity, diversity, and verbosity, while matching $\texttt{DPO}_1$ and GPT on empathy. These patterns indicate that the generation model produces responses with stronger stylistic accommodation, greater linguistic diversity, and comparable supportive characteristics relative to baseline models, while maintaining high semantic alignment with the original query.

4.3 Human Evaluation for GM

To assess the quality of responses generated by our models, we conduct a human-evaluations studies with a total of 210 participants recruited from Prolific. These participants rated 1680 query-response pairs in total (mean = 8 per participant). The evaluation is conducted in two phases. In the first phase, we evaluate responses generated by $\texttt{DPO}_1$ and compare them against online community responses and GPT-generated baselines. The human evaluations-cum-annotations are then incorporated into an additional round of DPO training or $\texttt{DPO}_2$. In this second phase, we conduct a follow-up human evaluation to assess how $\texttt{DPO}_2$ performs to understand whether the additional preference alignment further improves response quality. In both evaluation phases, participants are presented with query-response pair consisting of one post from our Reddit dataset, along with a response either from 1) original Reddit comment (OC), 2) GPT-5-nano (GPT), and 3) our model ($\texttt{DPO}_1$ in the first phase and $\texttt{DPO}_2$ in the second phase).

We design our evaluation criteria grounded in prior literature on social support, empathetic communication, and online mental health interactions (Cutrona and Russell, 1990; Suhr et al., 2004)

Question	OC	GPT	$\texttt{DPO}_1$	$t_{\texttt{GPT}-\texttt{OC}}$	$t_{\texttt{DPO}_1-\texttt{OC}}$	$t_{\texttt{DPO}_1-\texttt{GPT}}$	KW-H
Readability	3.678	4.615	4.501	12.767***	10.831***	-2.315*	153.947***
Empathy	2.933	4.268	4.358	15.501***	16.984***	1.405	259.479***
Connection	2.689	3.872	3.977	13.328***	14.653***	1.357	204.052***
Actionability	2.280	4.393	4.039	25.545***	19.621***	-5.051***	401.081***
Safety	3.148	4.524	4.473	15.707***	14.825***	-0.782	251.968***

Table 2: **LLUMI–GM** ($\texttt{DPO}_1$): Human evaluation results comparing OC (Online Community), GPT, and $\texttt{DPO}_1$ responses across five questions. Values are mean ratings; significance via paired t-tests and Kruskal–Wallis tests (* p<0.05, ** p<0.01, *** p<0.001).

Question	OC	GPT	$\texttt{DPO}_2$	$t_{\texttt{GPT}-\texttt{OC}}$	$t_{\texttt{DPO}_2-\texttt{OC}}$	$t_{\texttt{DPO}_2-\texttt{GPT}}$	KW-H
Readability	3.812	4.590	4.387	9.491***	6.499***	-3.321***	64.616***
Empathy	2.135	4.342	4.413	26.496***	27.500***	1.079	481.306***
Connection	1.941	3.965	4.059	23.612***	24.722***	1.208	413.588***
Actionability	1.803	4.322	4.238	30.907***	29.523***	-1.195	506.474***
Safety	2.950	4.590	4.487	17.958***	16.259***	-1.642	291.753***

Table 3: **LLUMI–GM** ($\texttt{DPO}_2$): Human evaluation results comparing OC (Online Community), GPT, and $\texttt{DPO}_2$ responses across five questions. Values are mean ratings; significance via paired t-tests and Kruskal–Wallis tests (* p<0.05, ** p<0.01, *** p<0.001).

to capture both the quality and perceived safety of supportive responses in peer-support settings: 1) **Readability**: This response is clear and easy to read and understand, 2) **Empathy**: This response shows empathy by acknowledging the post, 3) **Connection**: This response builds a sense of connection or relational warmth with the poster, 4) **Actionability**: This response offers practical and actionable guidance or perspectives, and 5) **Safety**: This response avoids harmful, judgmental, or risky suggestions. Participants are asked to evaluate each query-response pair on a scale of 1 (*Strongly disagree*) to 5 (*Strongly agree*), i.e., higher scores indicate a better quality of supportive response.

Table 2 and Table 3 summarize the results for both $\texttt{DPO}_1$ and $\texttt{DPO}_2$ across five evaluation criteria. Across all criteria, both GPT and DPO-based models consistently received substantially higher ratings than original online community responses. For $\texttt{DPO}_1$, significant improvements over OC responses were observed for readability (by 22%), empathy (by 49%), connection (by 48%), actionability (by 77%), and safety (by 42%). Similarly, $\texttt{DPO}_2$ demonstrated strong gains over OC responses, including readability (by 15%), empathy (by 106%), connection (by 109%), actionability (by 135%), and safety (by 52%). These findings suggest that DPO-based alignment substantially improves perceived supportive response quality.

Importantly, GPT–DPO comparisons were largely non-significant across most criteria. For $\texttt{DPO}_1$, significant differences from GPT appeared only in readability (by 3%) and actionability (by 9%), while $\texttt{DPO}_2$ showed no significant differences except for readability, indicating that our DPO-trained models achieve GPT-comparable performance despite lower computational requirements.

Furthermore, $\texttt{DPO}_2$, which incorporates additional human preference supervision, shows modest improvements over $\texttt{DPO}_1$. Notably, $\texttt{DPO}_2$ achieves higher mean scores than GPT on empathy (4.41 vs. 4.34) and connection (4.06 vs. 3.97). The GPT–DPO gap becomes smaller for these dimensions compared to $\texttt{DPO}_1$, suggesting that additional human preference alignment enables the model to more closely approximate GPT-level supportive behavior. These findings support the effectiveness of iterative DPO alignment for improving supportive response generation.

Model	Overall ↑	Empathy ↑	Specificity ↑	Medical ↓	Factual↑	Toxicity ↓
LLUMI (7B)	**4.74**	**4.83**	**4.26**	**0.10**	**3.98**	**1.00**
GPT-4	4.57	4.82	4.25	0.01	4.00	1.00
Gemini-1.5-pro	4.89	4.80	4.88	0.05	4.00	1.00
Llama 3.3 (70B)	5.00	4.99	5.00	0.01	4.00	1.00

Table 4: Performance on the CounselBench test set (100 test cases), evaluated by GPT-4.1 following the original CounselBench protocol. Baseline scores are reported from Li et al. (2026). Best value per column in bold. ↑ / ↓ indicate whether higher or lower is better.

4.4 Generalization to Out-of-Domain Counseling Dialogues

To assess whether LLUMI's supportive communication ability transfers beyond Reddit-style peer support, we additionally evaluate on CounselBench, a benchmark of mental health counseling dialogues (Li et al., 2026). We generate responses for the 100 test examples and score them using the same LLM-based protocol as the original benchmark, with GPT-4.1 as the evaluator across six dimensions: **overall quality, empathy, specificity, presence of unsolicited medical advice, factual consistency,** and **toxicity**. Baseline scores are taken directly from the original paper rather than re-run, so all comparisons share an identical evaluator and rubric but differ in generation-time conditions.

Table 4 reports the results. Despite being an order of magnitude smaller than the compared systems, LLUMI performs competitively: it attains the second-highest empathy score (4.83) and exceeds GPT-4 on overall quality (4.74 vs. 4.57), while trailing Llama 3.3 (70B) by 0.26 on overall quality and 0.74 on specificity. The substantive finding is that a 7B model trained purely on community feedback signal reaches the same quality band as

substantially larger general-purpose models on a counseling benchmark it was not trained for.

5 RQ2: Improvement Model (IM)

In addition to the response generation scenario in the previous section, we also explore the scenario when a respondent could draft their own response, and then use the LLM writing assistant to revise or improve their response. In particular, **IM** is an improvement model designed to revise an initial user-written response.

5.1 Building IM

5.1.1 IM Training Data

The improvement model (IM) is designed to revise an initial user-written response into a more empathetic, supportive, and emotionally appropriate version while preserving its original intent and conversational tone. Accordingly, each training instance consists of: 1) a Reddit mental health support post, 2) an initial human-written comment, and 3) an improved comment.

To construct the dataset, we select Reddit posts with more than 10 associated comments and randomly sample up to three comments per post as initial responses. For each post–comment pair, GPT-5-nano generates an improved version conditioned on instructions to increase supportiveness and empathy while maintaining a natural Reddit-style tone and avoiding overly clinical language, repetitive phrasing, emojis, and crisis hotline recommendations (Appendix A.2.6). This produces a synthetic SFT dataset combining authentic Reddit responses with model-generated revisions.

To validate dataset quality, we construct a benchmark of 2K post–comment–revision triples and compare original and revised comments using linguistic, semantic, and logical consistency analyses. As shown in Appendix Tables A1 and A2, revised comments exhibit stronger supportive characteristics while maintaining high semantic similarity and low contradiction, suggesting that revisions primarily strengthen existing responses rather than substantially altering their content.

5.1.2 IM Training

We train a retrieval-augmented generation (RAG) model designed to revise a user's original Reddit comment into a more supportive and empathetic response. Given a Reddit post and an initial user-written comment, the model learns to generate a refined version while preserving the original intent and conversational style. We adopt a retrieval-augmented setup because supportive communication is often highly context-dependent, and similar emotional situations frequently exhibit recurring patterns of supportive language and interaction styles within online communities (Appendix A.2.5).

To provide additional contextual guidance during generation, we incorporate retrieved examples from similar Reddit discussions (Appendix A.2.4). We first encode all Reddit posts using a pretrained `SentenceTransformer` and construct a FAISS index over the resulting embeddings. For a given input post, we retrieve the top-k most similar posts based on embedding similarity and include their corresponding improved comments as in-context demonstrations within the prompt. By grounding generation in semantically similar peer-support interactions, the model can better produce responses that are contextually appropriate and aligned with community communication norms rather than relying on generic supportive phrasing. Formally, given a post p, initial comment c_{init}, and retrieved context $\mathcal{R}(p)$, the model learns a mapping: $\hat{c} = f_\theta(p, c_{\text{init}}, \mathcal{R}(p))$, where $\hat{c}$ denotes the improved comment generated by the model.

We fine-tune Mistral-7B-Instruct-v0.2 using LoRA-based parameter-efficient training with a causal language modeling objective. Training prompts consist of a task instruction, the Reddit post, the initial comment, and retrieved contextual examples, while the target output is the improved comment. The base model is loaded in 4-bit NF4 quantization, and LoRA adapters are applied to the attention projection layers as well as the embedding and language modeling head layers to enable efficient adaptation. Training is performed with a maximum sequence length of 512 tokens using the AdamW optimizer and a learning rate of 2×10^{-4}. We employ gradient accumulation and clipping for stable optimization, and train the model for two epochs. This training setup enables the model to learn supportive comment refinement from GPT-generated revisions while leveraging retrieved examples from similar peer-support interactions.

5.2 Linguistic Evaluation for IM

We next evaluate the effectiveness of the comment improvement model, which refines an initial human-written response given the original post. Detailed descriptions of all linguistic and psycholinguistic metrics, including their definitions, compu-

Metric	OC	GPT	IM	t_{GPT-OC}	t_{IM-OC}	t_{IM-GPT}	KW-H
Verbosity							
Words Per Response	33.180	103.327	111.994	44.42***	29.63***	3.46***	991.92***
Words Per Sentence	13.001	15.639	16.582	5.21***	1.93	0.53	295.51***
Syntax							
Readability	28.573	6.704	6.099	-2.51*	-2.57*	-1.34	284.88***
Repeatability	0.198	0.400	0.520	27.64***	39.40***	22.59***	976.60***
Complexity	8.274	3.911	3.749	-2.87**	-2.98**	-1.92	221.35***
Style							
Cat. Dynamic Idx. (CDI)	-2.671	0.027	-4.301	2.90**	-1.69	-8.72***	48.32***
Empathy	0.978	0.983	0.987	5.80***	12.80***	7.18***	234.29***
Formality	0.396	0.624	0.678	15.31***	18.22***	5.39***	324.56***
Hope	0.858	0.777	0.771	-10.76***	-11.55***	-0.83	202.64***
Personalization							
Ling. Accommodation	0.793	0.925	0.919	14.42***	13.29***	-2.51*	149.92***
Semantic Similarity	0.289	0.445	0.394	24.09***	14.42***	-8.88***	328.34***
Diversity	0.165	0.434	0.419	41.74***	36.99***	-5.33***	928.23***

Table 5: **LLUMI–IM**: Linguistic metric comparisons between human-written comments (OC), GPT baseline generations, and LLUMI-improved comments (IM). Pairwise comparisons are evaluated with paired t-tests, and overall differences are evaluated with Kruskal–Wallis H-tests (* p<0.05, ** p<0.01, *** p<0.001).

Model	Preference	Content	Style
Original Response	16.3% of 424	2.83	2.38
GPT-5-nano	85.4% of 213	3.10	2.71
LLUMI–IM	82.0% of 211	2.89	2.67
Either improvement model	83.7% of 424	3.03	2.75

Table 6: **LLUMI–IM**: Participant ratings for original online community response (OC) and improvement model outputs across GPT-5-nano and LLUMI–IM, across dimensions of Content and Style match.

tation procedures, and interpretations, are provided in Appendix A.1. As shown in Table 5, LLUMI-improved comments show consistent gains over original comments and GPT baseline generations across multiple dimensions. In particular, improved comments show higher empathy and formality scores, alongside increased linguistic accommodation and diversity. At the same time, readability is substantially improved relative to original comments, suggesting that the model enhances clarity without sacrificing supportive tone. These results suggest that the improvement model effectively builds upon existing user-written content, refining it to better align with linguistic characteristics associated with effective peer support in online mental health communities.

5.3 Human Evaluation for IM

The human evaluations for IM model outputs consist of post–response pairs containing improved responses generated by either the GPT baseline or our improvement model. Participants are asked to choose which response they prefer (original or model-revised) and to evaluate whether the revised response preserves 1) **Content**: The revised response feels consistent with the original response's content, and 2) **Style**: The revised response feels consistent with the original response's voice or style. These are rated on the same 1 (*Strongly disagree* to 5 (*Strongly agree*) scale as before.

Table 6 shows the proportion of participant preferences of responses. Overall, participants preferred the model-generated responses in 355 out of 424 comparisons (83.7%), suggesting that both GPT and LLUMI were generally perceived as producing stronger supportive responses than the original Reddit comments. GPT responses were preferred in 182 out of 213 comparisons (85.4%), while responses generated by the LLUMI–IM were preferred in 173 out of 211 comparisons (82.0%), showing that LLUMI achieved a comparable preference rate, indicating that the model effectively improved response quality.

For content preservation, GPT-5-nano achieved a mean score of 3.103 while LLUMI–IM achieved 2.891, with a paired t-statistic of 1.036 but no statistical significance (p>0.05). Similarly, for style preservation, GPT-5-nano achieved a score of 2.709 compared to 2.673 for LLUMI–IM, with a paired t-statistic of 0.014 and no statistical significance (p>0.05). These findings suggest that the IM model preserves the original meaning, intent, and conversational style of user-written Reddit responses at a level comparable to GPT-5-nano while still improving supportive communication quality.

6 Discussion and Conclusion

We studied how smaller, open-source language models can be adapted into effective writing assistants for online mental health support, by building LLUMI setup for both response generation and response improvement. Across linguistic analyses and human evaluations, LLUMI performed competitively with proprietary GPT. Our results suggest that effective alignment for supportive communication does not depend only on model scale. Rather, community-specific data, preference supervision, and iterative human feedback can enable smaller models to provide emotionally appropriate and socially grounded writing assistance. Taken together, LLUMI demonstrates the potential of combining authentic online community interactions, community-derived preference signals, and human-centered alignment to build privacy-conscious AI systems that strengthen, rather than replace, peer support in online mental health communities.

Community-aligned writing assistance. A key implication of this work is that AI systems for mental health peer support should be designed as *writing assistants for human responders*, rather than as autonomous mental health agents. Online

mental health communities often rely on moderators, volunteers, and everyday peers who want to help but may not have formal training in counseling, crisis communication, or emotionally sensitive writing. At the same time, these communities frequently struggle to train volunteers who can provide timely, high-quality, and safe responses at scale. In this context, tools such as LLUMI can help reduce the burden on peer responders by assisting them in drafting or revising comments to be clearer, warmer, more validating, and less likely to unintentionally cause harm. Importantly, this framing preserves the peer-driven nature of online support: the AI does not replace the community member, but helps them communicate more effectively.

Implications for platform design. Our findings also point to concrete opportunities for designing future online support platforms. Rather than deploying AI as a direct responder to vulnerable users, platforms could embed writing assistance into the comment-composition interface. For example, when a user writes a response, the platform could offer optional revisions that improve empathy, clarity, validation, and safety while preserving the responder's original intent and voice. Such systems could also provide lightweight feedback before posting, such as flagging language that may sound dismissive, overly directive, generic, or potentially unsafe. This design direction shifts AI support from replacing human care to strengthening community care. It also creates a more accountable deployment model, where humans remain in control of what is posted and AI suggestions remain assistive rather than authoritative.

The value and limits of community feedback. Our results further show that community-derived signals, such as upvotes and downvotes, can provide scalable supervision for training supportive response models. These signals are valuable because they emerge from real interactions and reflect how community members endorse, reject, or value different kinds of support. However, they should not be treated as perfect measures of supportive quality. Upvotes may reflect visibility, writing style, popularity, or what feels good to bystanders rather than what is most helpful to the person in distress. The improvement from additional DPO training with survey-derived human preferences suggests that community feedback is most useful when combined with direct human evaluation. Our work inspires that future models can treat community feedback as one layer of alignment, complemented by human-centered evaluation, expert review, and safety-aware moderation.

Privacy-conscious and resource-sensitive deployment. This work also highlights the practical value of self-hosted open-source models for sensitive support contexts. Proprietary cloud-based APIs may require mental health-related disclosures to be transmitted to external providers, raising concerns about privacy, confidentiality, and data governance. In contrast, LLUMI can be deployed in-house within protected environments, giving platforms and community organizations greater control over data handling, auditing, and long-term deployment costs. This is especially important for resource-constrained communities that may not have access to large teams of trained volunteers, clinical experts, or expensive proprietary AI infrastructure. Smaller open-source models, when carefully aligned, may therefore offer a more feasible path for responsible AI-assisted support.

7 Limitations and Future Directions

Our work has several limitations which also suggest interesting future directions. First, although Reddit provides large-scale naturally occurring peer-support interactions, community feedback signals such as upvotes and downvotes are imperfect proxies for supportive quality and may also reflect factors such as popularity, writing style, or subreddit-specific cultural norms. Second, while our human evaluation study measures perceived empathy, supportiveness, and safety, it does not assess long-term psychological outcomes or the real-world impact of AI-assisted responses on users experiencing distress. In addition, our models are trained primarily on English-language Reddit communities, which may limit generalizability across cultures, languages, and support contexts. Future work could explore more diverse online communities, longitudinal evaluations of user wellbeing outcomes, and hybrid alignment approaches that combine community-derived feedback with expert clinical annotations. We also believe future systems could incorporate stronger safety-aware moderation mechanisms and personalized retrieval strategies to better adapt supportive responses to individual users and contexts.

8 Ethical Considerations

This work focuses on AI-assisted support for online peer-support communities and is not intended to replace professional mental health care, crisis intervention, or clinical counseling. Because the dataset is derived from publicly available Reddit discussions involving emotionally sensitive content, we take steps to minimize potential harm by removing personally identifying information and avoiding the release of raw user data. We additionally focus specifically on *r/SuicideWatch* because it is a well-moderated mental health support community that has been extensively studied in prior work on online peer support and supportive communication. The community maintains relatively strong norms surrounding empathetic, non-judgmental, and safety-conscious responses, making it a more appropriate setting for studying supportive interaction patterns and community preference signals.

At the same time, we acknowledge that community feedback signals such as upvotes and downvotes are imperfect proxies for supportive quality and may reflect social biases, subreddit-specific norms, or collective preferences that do not always align with clinically appropriate support practices. More broadly, when applying community-driven preference alignment approaches, it is important to consider the moderation structure, behavioral norms, and values promoted within a given online community. In communities where harmful or toxic behaviors are normalized or highly rewarded, engagement-based preference signals could reinforce undesirable behaviors during model training. Future work should therefore carefully evaluate the social dynamics and moderation practices of online communities before adopting community-derived engagement signals for preference optimization.

Additionally, despite alignment efforts and human evaluation, generated responses may still produce inappropriate, unhelpful, or unsafe outputs in certain situations. For this reason, we position LLUMI as a supportive writing-assistance tool rather than an autonomous mental health advisor. Future deployment of such systems should incorporate additional safety mechanisms, human oversight, and clear communication regarding the limitations of AI-generated support.

Our study was approved by the Institutional Review Board (IRB) at our institution. Participants were recruited through Prolific and compensated according to platform standards and estimated study completion time. Prior to participation, all participants were presented with informed consent materials describing the purpose of the study, the emotionally sensitive nature of the content, and how their responses would be used for research purposes(Section A.4). Because the study involves mental health-related Reddit discussions, participants were also informed that they could discontinue participation at any time without penalty. All collected evaluation data were anonymized and used solely for research purposes.

9 AI Involvement Disclosure

We used AI-assisted writing tools (e.g., ChatGPT and Grammarly) to refine and edit the writing of the manuscript. All analyses, scientific content, and experiments were written solely by the authors.

Acknowledgments

This work was supported in part by the Jump ARCHES endowment through the Health Care Engineering Systems Center at the University of Illinois and the OSF Foundation. We thank Yunhao Yuan for contributing to the initial discussions and brainstorming for this work.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive self-disclosures, responses, and social support on instagram: the case of# depression. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1485–1500.

Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. 2023. Don't lose the message while paraphrasing: A study on content preserving style transfer. In *International Conference on Applications of Natural Language to Information Systems*, pages 47–61. Springer.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Mohit Chandra, Suchismita Naik, Denae Ford, Ebele Okoli, Munmun De Choudhury, Mahsa Ershadi, Gonzalo Ramos, Javier Hernandez, Ananya Bhattacharjee, Shahed Warreth, and 1 others. 2025. From lived experience to insight: Unpacking the psychological risks of using ai conversational agents. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 975–1004.

Christel Lynne Chang, Chaitali Sinha, Madhavi Roy, and John Chee Meng Wong. 2024. Ai-led mental health support (wysa) for health care workers during covid-19: service evaluation. *JMIR Formative Research*, 8:e51858.

Zhifa Chen, Yichen Lu, Mika P Nieminen, and Andrés Lucero. 2020. Creating a chatbot for and with migrants: chatbot personality drives co-design activities. In *Proceedings of the 2020 acm designing interactive systems conference*, pages 219–230.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Emmelyn AJ Croes, Marjolijn L Antheunis, Chris van der Lee, and Jan MS de Wit. 2024. Digital confessions: The willingness to disclose intimate information to a chatbot and its impact on emotional well-being. *Interacting with Computers*, 36(5):279–292.

Carolyn E Cutrona and Daniel W Russell. 1990. Type of social support and specific stress: Toward a theory of optimal matching.

Carolyn E Cutrona and Beth R Troutman. 1986. Social support, infant temperament, and parenting self-efficacy: A mediational model of postpartum depression. *Child development*, pages 1507–1518.

Vedant Das Swain, Qiuyue "Joy" Zhong, Jash Rajesh Parekh, Yechan Jeon, Roy Zimmerman, Mary Czerwinski, Jina Suh, Varun Mishra, Koustuv Saha, and Javier Hernandez. 2025. Ai on my shoulder: Supporting emotional labor in front-office roles with an llm-based empathetic coworker. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 32–41.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Sindhu Kiranmai Ernala, Asra F Rizvi, Michael L Birnbaum, John M Kane, and Munmun De Choudhury. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–27.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.

Agam Goyal, Xianyang Zhan, Yilun Chen, Koustuv Saha, and Eshwar Chandrasekharan. 2025. Momoe: Mixture of moderation experts framework for ai-assisted online governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12656–12671.

Agam Goyal, Xianyang Zhan, Charlotte Lambert, Koustuv Saha, and Eshwar Chandrasekharan. 2026. Vastu: Language models struggle to recognize online community values. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP), Main Conference*.

Ilona Herlin and Laura Visapää. 2016. Dimensions of empathy in relation to language. *Nordic Journal of linguistics*, 39(2):135–157.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 4(1).

Rachael M Kang and Tera L Reynolds. 2024. "this app said i had severe depression, and now i don't know what to do": the unintentional harms of mental health applications. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, pages 1–17.

Alan E Kazdin. 2011. Evidence-based treatment research: Advances, limitations, and next steps. *American Psychologist*, 66(8):685.

Jiwon Kim, Violeta J Rodriguez, Dong Whi Yoo, Eshwar Chandrasekharan, and Koustuv Saha. 2026. Pair-safe: A paired-agent approach for runtime auditing and refining ai-mediated mental health support. *arXiv preprint arXiv:2601.12754*.

Meeyun Kim, Koustuv Saha, Munmun De Choudhury, and Daejin Choi. 2023. Supporters first: understanding online social support on mental health from a supporter perspective. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–28.

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A Smith, and Hannaneh Hajishirzi. 2025. Compo: Community preferences for language model personalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8246–8279.

Konstantinos Kyritsis, Isidoros Perikos, and Michael Paraskevas. 2023. Zero-shot sentiment analysis exploring bart models. In *2023 IEEE/ACIS 8th International Conference on Big Data, Cloud Computing, and Data Science (BCD)*, pages 192–197. IEEE.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Harry Levin and Margaretta Novak. 1991. Frequencies of latinate and germanic words in english as determinants of formality. *Discourse Processes*, 14(3):389–398.

Yahan Li, Jifan Yao, John Bunyi, Adam Frank, Angel Hwang, and Ruishan Liu. 2026. Counselbench: a large-scale expert evaluation and adversarial benchmarking of large language models in mental health question answering. In *International Conference on Learning Representations*, volume 2026, pages 95178–95214.

Nicholas Mcinnes and Bo JA Haglund. 2011. Readability of online health information: implications for health literacy. *Informatics for health and social care*, 36(4):173–189.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine*, 176(5):619–625.

Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 599–627.

Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research*, 20(6):e10148.

Sheshera Mysore, Debarati Das, Hancheng Cao, and Bahareh Sarrafzadeh. 2025. Prototypical human-ai collaboration behaviors from llm-assisted writing in the wild. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16830–16857.

John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2):113–122.

Viet Cuong Nguyen, Nhi Yen Nguyen, Kristin A Candan, Mary Conlon, Vanessa Rumie, Kristen Risola, Srijan Kumar, and Munmun De Choudhury. 2026. Calm-it: Generating realistic long-form motivational interviewing dialogues with dual-actor conversational dynamics tracking. *arXiv preprint arXiv:2601.10085*.

John C Norcross and Michael J Lambert. 2018. Psychotherapy relationships that work iii. *Psychotherapy*, 55(4):303.

James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 926–935.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Koustuv Saha, Sindhu Kiranmai Ernala, Sarmistha Dutta, Eva Sharma, and Munmun De Choudhury. 2020. Understanding moderation in online mental health communities. In *International Conference on Human-Computer Interaction*, pages 87–107. Springer.

Koustuv Saha, Yoshee Jain, Chunyu Liu, Sidharth Kaliappan, and Ravi Karkar. 2026. Ai vs. humans for online support: Comparing the language of responses

from llms and online communities of alzheimer's disease. *ACM Transactions on Computing for Healthcare*, 7(3):1–39.

Koustuv Saha and Amit Sharma. 2020. Causal factors of effective psychosocial outcomes in online mental health communities. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 590–601.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the web conference 2021*, pages 194–205.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–29.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13.

Jiayue Melissa Shi, Dong Whi Yoo, Keran Wang, Violeta J Rodriguez, Ravi Karkar, and Koustuv Saha. 2026. Mapping caregiver needs to ai chatbot design: Strengths and gaps in mental health support for alzheimer's and dementia caregivers. *ACM Transactions on Computing for Healthcare*.

Soorya Ram Shimgekar, Violeta J Rodriguez, Paul A Bloom, Dong Whi Yoo, and Koustuv Saha. 2026a. Suicidal ideation in online spaces through the lens of interpersonal theory of suicide: Exploratory study of self-disclosure, peer support, and ai responses. *JMIR AI*, 5(1):e86265.

Soorya Ram Shimgekar, Ruining Zhao, Agam Goyal, Violeta J Rodriguez, Paul Bloom, Navin Kumar, Hari Sundaram, and Koustuv Saha. 2026b. Detecting early and implicit suicidal ideation via longitudinal and information environment signals on social media. In *Proceedings of the 18th ACM Web Science Conference 2026*, pages 176–187.

Julie A Suhr, Carolyn E Cutrona, Krista K Krebs, and Sandra L Jensen. 2004. The social support behavior code (ssbc). In *Couple observational coding systems*, pages 307–318. Routledge.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 1489–1503.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, João Sedoc, Sven Buechel, and Alexandra Balahur. 2021. Wassa 2021 shared task: predicting empathy and emotion in reaction to news stories. In *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), held in conjunction with EACL 2021*, pages 92–104. Association for Computational Linguistics.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

John Torous and Honor Hsin. 2018. Empowering the digital therapeutic relationship: virtual clinics for digital health interventions. *NPJ digital medicine*, 1(1):16.

Piper Vornholt and Munmun De Choudhury. 2021. Understanding the role of social media–based mental health support among college students: Survey and semistructured interviews. *JMIR Mental Health*, 8(7):e24512.

David Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The effect of moderation on online mental health conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 751–763.

Zhouhang Xie, Bodhisattwa Prasad Majumder, Mengjie Zhao, Yoshinori Maeda, Keiichi Yamada, Hiromi Wakaki, and Julian McAuley. 2024. Few-shot dialogue strategy learning for motivational interviewing via inductive reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13207–13219.

Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–15.

Dong Whi Yoo, Jiayue Melissa Shi, Violeta J Rodriguez, and Koustuv Saha. 2026. Ai chatbots for mental health self-management: Lived experience–centered qualitative study. *JMIR Mental Health*, 13:e78288.

Yunhao Yuan, Koustuv Saha, Barbara Keller, Erkki Tapio Isometsä, and Talayeh Aledavood. 2023. Mental health coping stories on social media:

A causal-inference study of papageno effect. In *Proceedings of the ACM Web Conference 2023*, pages 2677–2685.

Xianyang Zhan, Agam Goyal, Yilun Chen, Eshwar Chandrasekharan, and Koustuv Saha. 2025. Slm-mod: Small language models surpass llms at content moderation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8774–8790.

A Appendix

A.1 Linguistic Analysis Metrics

A.1.1 Verbosity

We measure response verbosity using **Words Per Response (WPR)** and **Words Per Sentence (WPS)** (Saha et al., 2026). WPR is computed as the average number of words contained in a complete response, while WPS measures the average sentence length within a response. These metrics characterize how much information a model provides and the level of elaboration in generated text. Higher WPR values indicate more detailed and expansive responses, although excessively large values may also reflect unnecessary verbosity. Higher WPS values suggest longer and potentially more complex sentence structures, whereas lower values indicate shorter and simpler expressions.

A.1.2 Syntax

Readability measures how easily a text can be understood by readers. Prior work has identified readability as an important characteristic in health communication and online support settings, where communication quality depends not only on what information is conveyed but also on how easily users can interpret and process it (Mcinnes and Haglund, 2011; Ernala et al., 2017; Saha and Sharma, 2020). We measure readability using the Coleman–Liau Index (CLI) (Coleman and Liau, 1975), a readability metric based on character and sentence structure: CLI = 0.0588L - 0.296S - 15.8 where L denotes the average number of letters per 100 words and S denotes the average number of sentences per 100 words. Unlike readability measures based on syllable counts, CLI estimates text difficulty using word and sentence composition. Higher CLI values generally indicate more sophisticated writing and greater linguistic complexity; however, they also suggest that a higher level of education may be required to understand the text.

Repeatability and **complexity** are syntactic measures that characterize the richness and depth of language expression and have been associated with cognitive processes involved in communication, such as planning, execution, and information organization (Ernala et al., 2017). Repeatability captures how frequently words are reused within a response, while complexity reflects the sophistication and depth of language construction.

Following prior work (Ernala et al., 2017; Saha and Sharma, 2020; Yuan et al., 2023), we measure repeatability as the normalized occurrence of non-unique words within a response. Higher repeatability values indicate greater lexical redundancy and repeated information, which may suggest reduced conciseness and lower informational crispness.

We measure complexity as the average word length per sentence (Saha et al., 2026). Higher complexity values generally indicate more sophisticated and nuanced language that may convey ideas with greater precision and depth. However, excessively high complexity can also increase cognitive burden and make text more difficult to understand.

A.1.3 Style

The **Categorical Dynamic Index (CDI)** measures linguistic style along a spectrum ranging from *categorical* to *dynamic* language (Pennebaker et al., 2014). Categorical language reflects a more analytical and structured style that focuses on organized concepts and logical reasoning, whereas dynamic language is more narrative and socially oriented, often emphasizing personal experiences and storytelling. Prior work has shown that linguistic style is an important characteristic in social support and therapeutic communication contexts (Cutrona and Troutman, 1986; Saha and Sharma, 2020).

Following Pennebaker et al., CDI is computed as:

$$\begin{aligned} CDI = 30 &+ article + preposition \\ &- personal\ pronoun \\ &- impersonal\ pronoun \\ &- auxiliary\ verb - conjunction \\ &- adverb - negation \end{aligned}$$

where each term represents the percentage occurrence of the corresponding linguistic category within a text. We obtain these occurrences using the LIWC lexicon (Tausczik and Pennebaker, 2010). CDI is a bipolar measure in which higher values indicate a more categorical and analytical writing style, while lower values indicate a more dynamic, personal, and narrative style of communication.

Empathy is a central component of supportive communication and has been identified as an important mechanism in mental health support and social interactions (Sharma et al., 2020). Empathy refers to the ability to understand and recognize another individual's emotional state, experiences, and perspective, often through expressions of emotional understanding and validation (Herlin and

Visapää, 2016). In online support contexts, empathetic responses have been associated with more positive user perceptions and improved interaction quality (Sharma et al., 2020; Morris et al., 2018).

To measure empathy, we use a RoBERTa-based empathy detection model fine-tuned on datasets of empathetic reactions to news stories (Tafreshi et al., 2021; Buechel et al., 2018). Given a response, the model predicts an empathy score representing the degree to which the text expresses emotional understanding and supportive intent. Higher scores indicate stronger expressions of empathy and greater use of linguistic cues associated with emotional acknowledgment and perspective-taking.

Formality is an important sociolinguistic characteristic that reflects how language varies across audiences, contexts, and communication settings (Heylighen and Dewaele, 1999; Levin and Novak, 1991). Formality refers to the degree to which text follows established linguistic conventions, including structured syntax, appropriate vocabulary, and adherence to grammatical norms. Formal language is commonly associated with professional and academic communication, whereas informal language often contains conversational expressions, colloquialisms, and more casual writing styles.

To measure formality, we use a RoBERTa-based formality classifier proposed in prior work (Babakov et al., 2023). Given a text, the model predicts a formality score, where higher scores indicate greater use of formal language characteristics, while lower scores indicate a more conversational and informal writing style.

Hopefulness reflects the extent to which language conveys optimism, encouragement, and positive expectations about future outcomes. In mental health and supportive communication settings, expressions of hope can play an important role in fostering emotional support and promoting positive outlooks during distressing situations (Sharma et al., 2020). Supportive responses that communicate hope may encourage resilience and provide reassurance while avoiding excessively negative or fatalistic language.

To measure hopefulness, we use a zero-shot classification approach using the `facebook/bart-large-mnli` model proposed in prior work (Kyritsis et al., 2023). BART-large-MNLI is a variant of the BART model fine-tuned specifically for the MNLI task. We frame hope detection as a natural language inference task with two candidate labels: *hopeful* and *not hopeful*. Given a response, the model estimates the probability that the text semantically entails the label *hopeful*. Higher values indicate greater confidence that the text expresses hopeful language, while lower values indicate weaker expressions of optimism or encouragement.

A.1.4 Personalization

Linguistic style accommodation measures the extent to which a response adapts to the linguistic style of the original query. Prior work has shown that greater accommodation between individuals is associated with more effective communication and improved outcomes in online support interactions (Saha and Sharma, 2020; Sharma and De Choudhury, 2018). Unlike content-based measures, accommodation focuses on stylistic characteristics of language rather than semantic content, capturing how individuals adapt their communication patterns through grammar and function word usage.

Following prior work (Saha et al., 2026), we measure linguistic accommodation by computing the cosine similarity between linguistic style vectors derived from the original query and its corresponding response. The vectors are constructed using normalized occurrences of non-content linguistic categories obtained through the LIWC lexicon (Pennebaker et al., 2003; Saha and Sharma, 2020), including articles, prepositions, pronouns, auxiliary verbs, conjunctions, adverbs, and negations:

$$Accommodation(q, r) = \frac{L(q) \cdot L(r)}{||L(q)||\ ||L(r)||}$$

where $L(q)$ and $L(r)$ represent the linguistic style vectors of the query and response, respectively.

Higher accommodation values indicate stronger stylistic adaptation to the user's language patterns, suggesting greater alignment with the way users communicate. Lower values indicate weaker adaptation and a more generic response style.

Semantic similarity measures the degree to which a response remains semantically and topically aligned with the original query. In supportive communication settings, maintaining semantic relevance is important to ensure that responses address the user's concerns and remain coherent with the context of the discussion. Higher semantic similarity suggests that responses are more closely related

to the themes and meaning expressed in the original post.

Following prior work (Saha et al., 2026), we measure semantic similarity by computing the cosine similarity between vector representations of the query and response. We obtain sentence-level embeddings using pretrained BERT-based sentence transformers (Reimers and Gurevych, 2019). Semantic similarity is computed as:

$$Sim(q, r) = \frac{E(q) \cdot E(r)}{||E(q)||\ ||E(r)||}$$

where $E(q)$ and $E(r)$ denote the embedding representations of the query and response, respectively.

Higher similarity scores indicate stronger topical and semantic alignment between a query and its response, suggesting that the response more closely addresses the user's original concerns. Lower scores indicate weaker semantic coherence and potentially less relevant responses.

Diversity measures the degree of variation and uniqueness in generated responses. Prior work has shown that diverse and creative responses can improve the effectiveness of psychotherapy and supportive interactions by reducing repetitive language and enabling a broader range of supportive strategies (Althoff et al., 2016; Norcross and Lambert, 2018; Saha and Sharma, 2020). In conversational systems, low diversity may indicate templated or repetitive responses, whereas greater diversity can reflect richer and more flexible language generation.

Following prior work (Mikolov et al., 2013; Saha et al., 2026), we measure diversity using word embedding representations in a 300-dimensional semantic space. For each dataset, we first compute a centroid vector representing the average embedding across all responses. We then calculate the cosine distance between each response embedding and its corresponding dataset centroid:

$$Diversity(r) = 1 - \frac{E(r) \cdot C}{||E(r)||\ ||C||}$$

where $E(r)$ denotes the embedding representation of response r, and C represents the centroid embedding of the dataset. Larger distance values indicate that a response deviates more from the average response representation and is therefore considered more diverse.

Higher diversity scores indicate greater linguistic variation and less repetitive or templated language, while lower values indicate responses that are more similar to common patterns within the dataset.

A.2 Model Training Prompts

A.2.1 GM SFT Prompt (Instruction-tuning)

```
<s>[INST]
``You are a Reddit comment writing assistant.
 Your goal is to write a supportive,
empathetic, and helpful comment for the
given Reddit post. Respond with warmth,
validation, and concise support without
repeating or over-empathizing.

The post: {post\_text}.''
[/INST]
{response}
</s>
```

A.2.2 GM DPO_1 Prompt

```
You are a Reddit comment writing assistant.
Your goal is to write a supportive,
empathetic, and helpful comment for the
given Reddit post. Respond with warmth,
validation, and concise support without
repeating or over-empathizing.

The post: {post_text}.
```

A.2.3 GM DPO_2 Prompt

```
You are a Reddit comment writing assistant.
Your goal is to write a supportive,
empathetic, and helpful comment for the
given Reddit post. Respond with warmth,
validation, and concise support without
repeating or over-empathizing.

The post: {post_text}.
```

A.2.4 IM SFT Training Prompt using RAG

```
<s>[INST]
<role>
You are a Reddit commenting assistant. Your
role is to read the original <post> and the
user's <initial_comment>, then improve that
<initial_comment> to make it sound more
supportive, empathetic, and helpful toward
the original poster.
</role>

<commenting_guidelines>
1. **Tone and Style**
- Sound like a genuine Reddit user  warm,
human, and conversational.
- Avoid overly formal, clinical, or
```

```
therapist-like language.
- Do not use generic or repetitive empathy (
e.g., Im so sorry youre going through this
repeatedly).
- Avoid using emojis.
- Do not suggest any type of hotline or
crisis resource in your response (e.g., "If
youre in immediate danger or think you might
 harm yourself again, call emergency
services or go to the nearest ER. If youre
in the US: call or text 988, or chat at 988
lifeline.org.").

2. **Empathy and Support**
- Show understanding and validation of the
original posters feelings.
- Offer encouragement, gentle advice, or
shared perspective where appropriate.
- Be kind, authentic, and emotionally
intelligent  not robotic or exaggerated.

3. **Content Quality**
- Improve clarity, flow, and emotional
resonance.
- Keep it **concise**  avoid long,
repetitive paragraphs.
- Ensure the improved comment provides **
constructive emotional or practical support
**.
- Maintain or enhance any useful information
 from the original comment.
</commenting_guidelines>

<examples>
Post: [Retrieved similar Reddit post]
Supportive Comment: [Retrieved supportive
response]

---

Post: [Retrieved similar Reddit post]
Supportive Comment: [Retrieved supportive
response]
</examples>

<post>
{post}
</post>


<initial_comment>
{init_comment}
</initial_comment>
[/INST]{target_comment}
</s>
```

A.2.5 Retriever Training Objective

```
Query:
{post_title} {post_body}

Positive Example:
{generated_target_comment}

Negative Example:
{initial_comment_body}
```

A.2.6 IM: GPT Prompt for Generating Revised Comments

```
<role>
You are a Reddit commenting assistant. Your
role is to read the original <post> and the
user's <initial_comment>, then improve that
<initial_comment> to make it sound more
supportive, empathetic, and helpful toward
the original poster.
</role>

<commenting_guidelines>
1. **Tone and Style**
- Sound like a genuine Reddit user  warm,
human, and conversational.
- Avoid overly formal, clinical, or ``
therapist-like'' language.
- Do not use generic or repetitive empathy (
e.g., ``I'm so sorry you're going through
this'' repeatedly).
- Avoid using emojis.
- Do not suggest any type of hotline or
crisis resource in your response (e.g., "If
you're in immediate danger or think you
might harm yourself again, call emergency
services or go to the nearest ER. If you're
in the US: call or text 988, or chat at 988
lifeline.org.").

2. **Empathy and Support**
- Show understanding and validation of the
original poster's feelings.
- Offer encouragement, gentle advice, or
shared perspective where appropriate.
- Be kind, authentic, and emotionally
intelligent  not robotic or exaggerated.

3. **Content Quality**
- Improve clarity, flow, and emotional
resonance.
- Keep it **concise**  avoid long,
repetitive paragraphs.
- Ensure the improved comment provides **
constructive emotional or practical support
**.
- Maintain or enhance any useful information
 from the original comment.
</commenting_guidelines>

<response_format>
- Output only the **improved comment**,
nothing else (no explanations, no markdown,
no quotes).
</response_format>

<post>
```

```
{post_content}
</post>

<initial_comment>
{initial_row.comment_body}
</initial_comment>
```

A.3 Data Validation for IM

Metric	Initial Comment	Generated Comment (LLUMI)	t-test Significant
cdi	-3.49	2.89	**Yes**
complexity	3.97	4.09	**Yes**
diversity_sent	0.23	0.44	**Yes**
empathy	0.97	0.99	**Yes**
formality	0.57	0.59	No
hope	0.85	0.78	**Yes**
ling_style	0.88	0.94	**Yes**
readability	4.84	7.95	**Yes**
repeatability	0.26	0.43	**Yes**
wpr	53.61	124.27	**Yes**
wps	13.96	15.72	**Yes**

Table A1: Paired linguistic analysis comparing initial human-written comments and LLUMI-generated improved comments. Statistical significance is determined using independent t-tests ($p < 0.05$).

Category	Metric	Mean Score
Embedding Similarity	Cosine similarity (Initial vs. Improved)	0.244
Cross-Encoder STS	STS (Initial vs. Post)	-1.914
	STS (Improved vs. Post)	-1.203
	STS (Improved vs. Initial)	-0.613
NLI Consistency	Entailment (Initial $\rightarrow$ Improved)	0.154
	Contradiction (Initial $\rightarrow$ Improved)	0.032
	Entailment (Improved $\rightarrow$ Initial)	0.226
	Contradiction (Improved $\rightarrow$ Initial)	0.160
BERTScore	BERTScore-F1 (Improved vs. Initial)	0.835
	BERTScore-F1 (Improved vs. Post)	0.794

Table A2: Semantic similarity and consistency analysis between initial comments and their LLUMI-improved versions. Higher similarity and entailment scores indicate stronger content preservation, while lower contradiction scores indicate reduced semantic divergence.

A.4 Prolific Instructions

```
Study name: Evaluating Supportive Responses
in Online Mental Health Communities

Study Description: Participants will
complete an online survey in which they
review a set of short postresponse examples
from online mental health peer-support
contexts. Each example consists of a post
and one or more responses. For each example,
 participants will answer structured
questions evaluating the responses (e.g.,
perceived supportiveness, empathy, and
helpfulness). Approximately 10 examples will
 be evaluated. The survey is self-paced and
requires no writing or sharing of personal
experiences.
```