

Assess and Prompt: A Generative RL Framework for Improving Engagement in Online Mental Health Communities

Bhagesh Gaur¹, Karan Gupta¹, Aseem Srivastava¹, Manish Gupta², Md. Shad Akhtar¹

¹IIIT Delhi, India; ²Microsoft, India

{bhagesh20558, karan21258, aseems}@iiitd.ac.in,
gmanish@microsoft.com, shad.akhtar@iiitd.ac.in

Abstract

Online Mental Health Communities (OMHCs) provide crucial peer and expert support, yet many posts remain unanswered due to missing support attributes that signal the need for help. We present a novel framework that identifies these gaps and prompts users to enrich their posts, thereby improving engagement. To support this, we introduce REDDME, a new dataset of 4,760 posts from mental health subreddits annotated for the span and intensity of three key support attributes: event (*what happened?*), effect (*what did the user experience?*), and requirement (*what support they need?*). Next, we devise a hierarchical taxonomy, CUETAXO, of support attributes for controlled question generation. Further, we propose MH-COPILOT, a reinforcement learning-based system that integrates (a) contextual attribute-span identification, (b) support attribute intensity classification, (c) controlled question generation via a hierarchical taxonomy, and (d) a verifier for reward modeling. Our model dynamically assesses posts for the presence/absence of support attributes, and generates targeted prompts to elicit missing information. Empirical results across four notable language models demonstrate significant improvements in attribute elicitation and user engagement. A human evaluation further validates the model’s effectiveness in real-world OMHC settings.

1 Introduction

The proliferation of Online Mental Health Communities (OMHCs) has evolved the landscape of the mental health support ecosystem, providing platforms where individuals can share experiences, seek advice, and obtain suggestions or support. These digital arenas have democratized access to mental health resources, especially for those who might otherwise face barriers to traditional support systems, possibly due to stigma, cost, or scarcity of experts (Naslund et al., 2016). Despite their

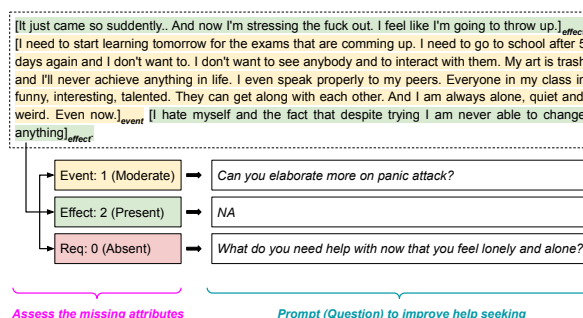


Figure 1: A sample post on a mental health subreddit. Our proposed framework attempts to identify support attributes and then enforce the language model to generate guiding questions.

potential, a significant challenge persists: many support-seeking posts remain unanswered, leaving individuals without the assistance they seek due to poor readability. Several times, user forget to mention essential details of what happened, what did the user experience and what support they need. How do we help users in better articulation of posts to increase the chance of receiving a response?

Previous studies have primarily focused on either assessing community engagement (Pruksachatkun et al., 2019; Sharma et al., 2020a; Kim et al., 2023) or automating responses (Sharma et al., 2022). At the same time, a rich body of prior work highlights Reddit’s prominence as a valuable resource for mental health research (Chen et al., 2023; Garg et al., 2023; Gaur et al., 2018; Rai et al., 2024; Lokala et al., 2022). While language models (LMs) have enhanced support providers’ experiences in empathy induction (Sharma et al., 2021; Srivastava et al., 2025b), a critical gap remains in directly assisting support seekers. Drawing inspiration from traditional therapeutic stages, building bond, identifying root causes, evaluating consequences, and clarifying goals (Adhikary et al., 2024; Benjet et al., 2016), our work translates these paradigms into online settings using state-of-the-art LMs. With better contextual understanding, interpersonal in-

sights (Romanowski et al., 2023), and explainable AI (Ibrahimov et al., 2024), we aim to bridge the gap by focusing on improving support seekers’ experiences in OMHCs.

To address these gaps systematically, we first curate a novel first-of-its-kind dataset, REDDME, extending the publicly available mental health corpus – BeCOPE (Srivastava et al., 2025a). REDDME contains 4760 posts, where each post is annotated for the spans and presence/absence of three pivotal support seeker’s cues: *event*, *effect*, and *requirement*. Next, we also propose a dedicated taxonomy of support attributes, CUETAXO, that allows controlled prompting for LMs and is responsible for question generation. Finally, we propose MH-COPILOT, a reinforcement learning (RL)-based framework tailored to enhance the posting behavior of support seekers on OMHCs. MH-COPILOT operates on two principles: (a) assessing the support-seeker’s post for missing support attributes and (b) prompting a question to better articulate their posts by using CUETAXO. Figure 1 illustrates an example of how a user can be prompted for different support seekers’ cue intensity levels. A reward model ensures the quality and relevance of these generations through RL.

We benchmark MH-COPILOT against four LLMs: Llama-3 (Grattafiori et al., 2024), Mistral (Jiang et al., 2023), Phi-3 (Abdin et al., 2024), and Gemma-2 (Riviere et al., 2024). Our proposed model demonstrates consistent improvements across all baselines, with a notable improvement of +27.49%, +17.80%, +2.81% and +17.54% on ROUGE-L, BLEU-4, BERTScore and METEOR respectively. We further conduct human evaluation and qualitative analysis to assess the quality of generations. Our contributions are summarized below:

- We propose a **novel problem** of enhancing support-seekers’ posting behavior using linguistic and contextual cues to identify missing attributes, prompting users to share more.
- A **novel dataset** containing mental health-related posts from Reddit, marked with key support attributes at span and intensity levels.
- A **reinforcement-learning-based framework** to assess support attributes and generate taxonomy-guided questions to enhance post responsiveness.
- Evaluation against notable LMs, where our approach consistently outperformed on the complete suite of evaluation metrics.

Reproducibility. The dataset and code for MH-COPILOT¹ is open sourced for research purposes.

2 Related Work

We study the support-seeking behavior on OMHCs. The study of psychological aspects of human behaviour and counselling dates long back (Weizenbaum, 1966). We aim to study these aspects for controlled question generation.

Online Mental Health Communities (OMHCs) and Support Provision.

The increasing reliance on digital platforms for mental health support highlights the need to enhance user experience in OMHCs. Kim et al. (2023) and Althoff et al. (2016) developed tools to improve interactions between support seekers and peer supporters. Despite progress, direct assistance for support seekers remains limited, a gap our work addresses. Studies have documented barriers to offline care (Olsson, 2016) and stigma in help-seeking (White and Dorman, 2001), underscoring the critical role of peer support. While Sharma et al. (2021) and Yang et al. (2019) have advanced empathic communication, others like Tanana et al. (2019) and Zhang et al. (2019) have worked on moderation and engagement. In contrast, our research focuses on enhancing responsiveness to support-seeking posts through controlled prompting.

Controlled Prompting in LLMs. Controlled prompting is essential for refining outputs from LLMs in precision-sensitive contexts (Evuru et al., 2024). While LLMs handle vast data, studies show LLMs often struggle with systematic generalization, producing content that is both coherent and high-quality (Petruzzellis et al., 2024). Output quality can be further improved using prompt optimization techniques (Kumar et al., 2024).

Reward Modeling and Self-Improvement in Language Modeling.

In sensitive domains like OMHCs, accuracy and appropriateness are vital. To gain more control, we incorporate self-improvement techniques (Zelikman et al., 2022; Srivastava et al., 2023) that refine model outputs during fine-tuning. We also use Direct Preference Optimization (DPO) (Rafailov et al., 2023) to guide the model toward generating more context-sensitive questions. This is particularly important when handling deviations from standard inputs

¹<https://github.com/flamenlp/MH-COPILOT>

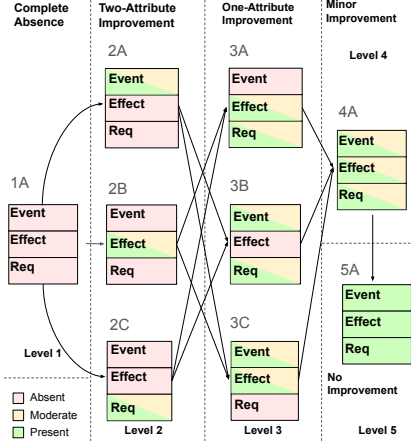


Figure 2: Taxonomy: CUETAXO.

	Event	Effect	Requirement
1A	Can you tell me what happened? You can be as specific as you like.	Could you describe the specific effect the event has had on you?	What kind of support or help you feel would be most beneficial?
2A	Can you elaborate more on X?	How did X make you feel?	What do you need help with now that X?
2B	What made you feel X?	Can you elaborate more on X?	What can help you overcome X?
2C	What happened that you want X?	Why are you wanting X?	Can you elaborate more on X?
3A	What made you feel X?	Can you elaborate more on X?	Can you elaborate more on X?
3B	What happened that you want X?	How did X make you feel?	Can you elaborate more on X?
3C	Can you elaborate more on X?	Why are you wanting X?	Can you elaborate more on X?
3C	Can you elaborate more on X?	What caused you to need X?	What do you need help with now that X?
4A	Can you elaborate more on X?	Can you elaborate more on X?	What can help you overcome X?
4A	Can you elaborate more on X?	Can you elaborate more on X?	Can you elaborate more on X?

Table 1: Taxonomy-based question prompts annotated for each taxonomy level. X signifies a user-mentioned entity. Questions are generated only for moderate or absent intensities; hence level 5A is not shown.

(Hosseini et al., 2024). By integrating controlled prompting and reward modeling, our framework enhances output quality while aligning better with the needs of mental health support seekers, addressing a key research gap.

3 Dataset: REDDME

We propose REDDME, a manually annotated corpus extended from the publicly available mental health subreddit corpus, BeCOPE (Srivastava et al., 2025a). BeCOPE categorizes posts into three primary categories: (a) *interactive*, if there are back-and-forth conversations between the OP (original poster) and peers (b) *non-interactive*, if the post engages peers, but the OP does not reply to peers (c) *isolated*, if the post does not have any comments. We select 4,760 posts and manually annotate them with support attributes: *event*, *effect*, and *requirement* on span (rationale) and intensity level.

Motivation for Attribute Selection. During a conversation with support seekers, therapists often provide support in stages over time (Ciletti, 2025). First, they build rapport with the patient, which is essential for creating a safe and trusting environment (Adhikary et al., 2024). Following this, they try to identify the root cause of the patient’s condition, which could be a single **event** or a series of events, potentially involving other individuals (Kirkbride et al., 2024; Cherry and Susman, 2023). Next, the therapist evaluates the potential **effects** of these events on the patient, which can range from depression and suicidal thoughts to changes in appetite or sleep (Benjet et al., 2016). It’s important to note that this evaluation is subjective, and the effects may not always be apparent (Tiwari and

Deshpande, 2020). Finally, the therapist identifies what the patient seeks (**requirement**) from the therapy, which helps keep them proactively involved and committed (Stewart et al., 2024). It’s crucial for understanding the support seeker’s perspective (Geurtzen et al., 2020; Saxena et al., 2022) before initiating therapy. Figure 3 shows an increase in the number of comments received by the support seeker as the level of each attribute increases.

Support Attributes. Building up on these theories from previous work, we define our three primary support attributes, which are critical to capture the help-seeking nature of posts: *event*, *effect*, and *requirement*.

- **Event** encapsulates the specific situation, activity, or event that is the focal point of the support seeker’s concern. The explicit detailing of such events provides a contextual background essential for overall background understanding, as suggested by Sharma et al. (2020b).
- **Effect** targets the impact or consequences of the identified event from the support seeker. By elucidating the effect, the post conveys the emotional or practical repercussions of the situation, thereby inviting more targeted and empathetic responses.
- **Requirement** lays out the expectation (e.g., informational support, instrumental aid, etc.) of support seeker from peers. It is crucial in directing the nature of the assistance sought, and thereby guiding the potential support trajectory. This aligns with the insights laid out by Sharma et al. (2022), highlighting the importance of clearly articulated needs for effective support.

Table 2 presents example text spans from Reddit

Attributes	Definition
Event	Encapsulates the specific situation, activity, or event that is the focal point of the support seeker’s concern. The explicit detailing of such events provides a contextual background essential for overall background understanding, as suggested by Sharma et al. (2020b). Example: “So I have been taking adderall 10mg for a little over a week and I’ve been testing it. Today I decided to take Adderall and coffee at the same time.”
Effect	Targets the impact or consequences of the identified event from the support seeker. By elucidating the effect, the post conveys the emotional or practical repercussions of the situation, thereby inviting more targeted and empathetic responses. Example: “I think this was a bad idea because I felt extremely jittery and anxious, but the worst part was I came down from them at the same time.”
Requirement	Lays out the expectation (e.g., informational support, instrumental aid, etc.) of support seeker from peers. It is crucial in directing the nature of the assistance sought, and thereby guiding the potential support trajectory. This aligns with the insights laid out by Sharma et al. (2022), highlighting the importance of clearly articulated needs for effective support. Example: “So if you guys got any tips that would be helpful, like if its the flu, or the Adderall, or caffeine, or anyone else has experienced this.”

Table 2: Definition and examples of support attributes.

posts illustrating each support attribute.

Annotation Process. The overall annotation process is crafted on three major fronts: (a) rationale (span-level) annotation considering the three support attributes, (b) intensity level annotation, and (c) taxonomy-based prompt question annotation.

Rationale Annotation. For each support attribute, we annotate spans corresponding to the event, effect, and requirement attributes by bounding them with special $\langle start \rangle$ and $\langle end \rangle$ tokens – we use $\langle es \rangle$, $\langle ee \rangle$ for event, $\langle efs \rangle$, $\langle efe \rangle$ for effect, and $\langle rs \rangle$, $\langle re \rangle$ for requirement, respectively. Due to the informal nature of the Reddit posts (incorrect grammar, inconsistent punctuation, absence of sentence markers, etc.), we execute the rationale annotation on a token level, with a preference for marking the complete sentences as much as possible. We share additional details pertaining to the rationale in Appendix A.

Intensity Annotation. To quantify the presence of support attributes, we employ a three-scale Likert rating system for annotation: *absent* (0), *moderately-present* or ‘*moderate*’ (1), and *well-described* or ‘*present*’ (2), indicating the degree to which each attribute is reflected in the original post.

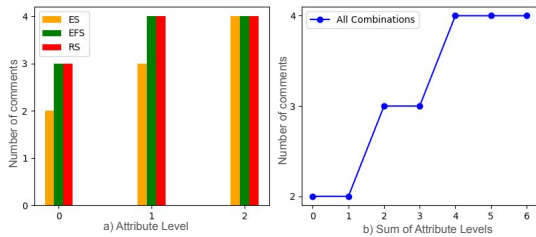


Figure 3: (a) Distribution of median number of comments for each level of the support attribute, (b) Median Number of comments for sum of level of event effect and requirement for a post.

For instance, if a post does not carry any *event* rationale, the corresponding intensity label is marked as *absent*. An example is shown in Figure 1.

Taxonomy-based Reference Question Annotation. We introduce a custom taxonomy, CUE-TAXO, that enables a structured and interpretable approach to prompting users with targeted questions for improving support attributes. As shown in Figure 2, CUE-TAXO consists of five levels, each representing varying degrees of completeness for the three support attributes within a post. Level 1 (none present; red-colored), Level 5 (all well-described; green-colored), and Levels 2-4 capturing partial expression (yellow-colored). To make our taxonomy generalizable, we have similar questions for the same intensities. For cases with absent attribute information, our evaluations reveal the need for customized questions that utilize context from other attributes in the follow-up suggestions for the support seekers. As a result, CUE-TAXO assists in the selection of the ideal template question, as depicted in Table 1. For instance, in the case where *event* is missing, we ask annotators to complete: *What happened that you wanted X?*. *X* is a placeholder that annotators can fill in with the help of rationales.

Annotator Details. We employ two annotators, aged 21–24, with expertise in linguistics and the relevant domain, to annotate the dataset. Additionally, an expert reviewer acts as a moderator and periodically evaluates the annotation quality. Both annotators are provided with guidelines and examples to support the consistent identification and categorization of each annotation type within the posts, which we present in the Appendix B. During the initial training phase, annotators participate in regular calibration sessions and discussions to align their interpretation of the guidelines and resolve any annotation discrepancies. Following

Split	#Posts	#Prompts	APoL	Support Attributes											
				Event			AEvL	Effect			AEfL	Requirement			AReL
				(words)	Ab	Mo		Pr	(words)	Ab		Mo	Pr	(words)	
Train Val Test	3331	5533	180.99	938	568	1825	66.35	1531	415	1385	26.50	1722	359	1250	19.62
	953	1583	168.25	268	162	523	61.31	438	118	397	26.72	493	104	356	18.51
	476	793	192.81	134	82	260	69.51	221	58	197	25.89	246	52	178	18.82
Total	4760	7909	179.62	1340	812	2608	65.70	2190	591	1979	26.48	2461	515	1784	19.31

Table 3: Statistical Analysis of REDDME. Train, val, test split is 70:20:10. **APoL**, **AEvL**, **AEfL**, and **AReL** are the average lengths of *post*, *event span*, *effect span*, and *requirement span*, respectively. **#Prompts** defines number of question-prompts. **Ab**: Absent, **Mo**: Moderately Present, **Pr**: Present.

the training phase, the annotators independently annotated two equally divided subsets of the dataset. To assess inter-annotator agreement, we calculated Cohen’s Kappa scores on a random sample of 50 instances. The resulting agreement scores were 0.885 for event, 0.886 for effect, and 1.000 for requirement, indicating a high degree of consistency.

Detailed statistics of REDDME are shown in Table 3. Further dataset details are in Appendix C.

4 Methodology

In this work, we propose MH-COPILOT, a novel framework that prompts support seekers with guiding questions based on their posts to elicit improvements. Our framework first identifies the presence of such attributes and further generates the required prompts via a hierarchical taxonomy, CUETAXO. To achieve this, we divide MH-COPILOT into four fundamental modules: (a) a contextual attribute-span identifier (CSpan), (b) a support attribute intensity classifier, (c) a guiding question generator, and (d) a verifier. A schematic diagram of framework is shown in Figure 4. Next, we discuss each of these modules in detail.

Contextual Span Identifier (CSpan). The CSpan module is responsible for identifying and extracting the relevant contextual spans that represent key attributes of a support-seeking post: *event*, *effect*, and *requirement*. These components provide essential context that can help responders better understand the situation described by the support seeker. We frame this as an entity extraction task and fine-tune RoBERTa (Liu et al., 2019) on the REDDME dataset to extract support attribute spans. Precisely, for each post P in the dataset, the input consists of tokens $\{t_1, t_2, \dots, t_n\}$ and corresponding annotated labels $\{a_1, a_2, \dots, a_n\}$, where each label could be one of event, effect or requirement. We utilize these spans further in the identification of the relevant attribute’s intensity.

Support Attribute Intensity Classifier. After contextual span identification, the next step is to assess the intensity of each attribute, as a typical multi-class, multi-label classification problem. The classifier takes the token-level span outputs from the CSpan module as input and yields an intensity vector, $V = \{v_{\text{event}}, v_{\text{effect}}, v_{\text{req}}\}$, where each $v_i \in \{0, 1, 2\}$ signifies absent (0), moderate (1), or present (2), respectively. Subsequently, these values act as the degree to which the support attributes are articulated in the post. We fine-tune RoBERTa for this task, which outperforms three baselines (Table 5). The output of this module guides the question-generator module by providing information on how well the attributes are expressed using our taxonomy.

Guiding Question Generator. We embed the CUETAXO taxonomy in the language model via prompting (Appendix D), by adding the levels of all the attributes in the prompt for the language model, enabling it to generate questions that are informed by the identified spans and their corresponding intensity levels. It ensures that generated prompts abide by the support attributes. We perform supervised fine-tuning of LLMs on REDDME to train *question generator* to learn the patterns of generating helpful prompts that align with the taxonomy. Given our prompt, we expect the output to be contextually relevant questions for each attribute. For example, if the event component is moderately described, the prompt generator generates a question such as, “Can you describe more about the event that led to your current feelings?”. Next, we utilize a reward model to control the generation via DPO (Rafailov et al., 2023).

Verifier and Reward Modeling. Verifier module validates the outputs of the question generator module across these dimensions:

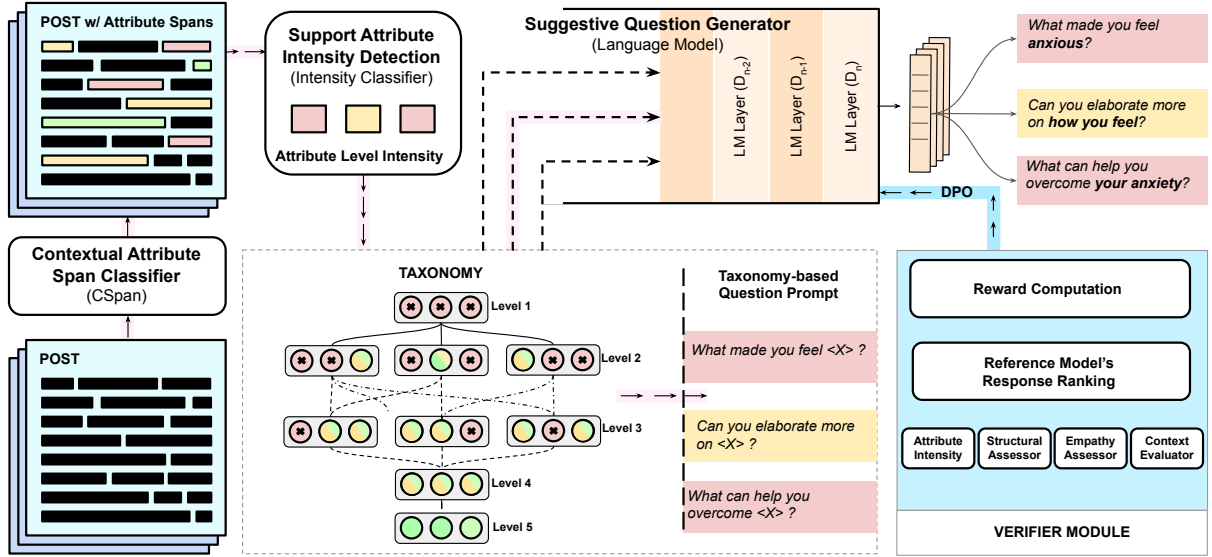


Figure 4: MH-COPILOT Framework: The four contributory modules are (a) contextual attribute span classifier, (b) support attribute intensity detection, (c) taxonomy-based question generation, and (d) verifier.

- **Support Attribute Category Classifier.** This multi-class classifier verifies whether each generated question aligns with the intended support attribute. We employ the best-performing baseline, RoBERTa, as a reference model in the verifier model. If the question correctly matches the target attribute, it receives a reward score of 1; otherwise, it receives a reward score of 0.
- **Contextual Grounding Evaluator.** The context map component uses a reference language model to evaluate how closely the generated question aligns with the original post. Considering the large and complex contexts, we employ Llama-3-70b (Grattafiori et al., 2024) for this purpose. A score between 0 and 1 is assigned, with higher scores indicating stronger contextual relevance.
- **Structural Adherence Assessor.** We ensure that the generations adhere to a pre-defined structure, defined within the taxonomy template. A score of 1 is for correct format and 0 for deviations.
- **Empathy Assessor.** We ensure that the generations are empathetic toward the support seeker, and carefully avoid triggering language or tone. We use Llama-3-70b to generate a score between 0 and 1; higher scores indicate higher empathy.

Reward Modeling. The final reward (r) for each sample is calculated by combining the scores from all reward parameters, as shown in Equation 1.

$$r = \sum_{i \in [1,2,3]} r_{Q_i}^{CC} \cdot r_{Q_i}^{CG} \cdot r_{Q_i}^{EA} \cdot r_{SA} \quad (1)$$

where Q_1, Q_2, Q_3 represent the questions for event, effect, and requirement components, and $r^{CC}, r^{CG}, r^{EA}, r^{SA}$ are the scores for category correctness using attribute classifier, contextual grounding using context map, empathy accessor and structure adherence using structure assessor, respectively. We augment preference data to train MH-COPILOT with DPO by generating two responses from variant with taxonomy enabled with supervised fine-tuning (SFT+CUETAXO) for the training dataset. These responses are then passed to the verifier module to get a reward score for each of them, r_1 and r_2 using Eq. 1.

Response with the higher and lower rewards are chosen as preferred (y_p) and non-preferred (y_{np}) responses for DPO respectively. This ranking contains the nuances of the context scores, structure scores, and empathy scores. We create the preferential dataset for the responses as: $\mathcal{D} = \{x^{(i)}, y_p^{(i)}, y_{np}^{(i)}\}$. Here, $x^{(i)}$ is the prompt for generating $\langle y_p^{(i)}, y_{np}^{(i)} \rangle$. The resulting preferential dataset is utilized for reward tuning using the likelihood function shown in Eq. 2. Hinge-loss is used as shown in Eq. 3, where σ denotes the logistic function, and β is a scaling parameter.

$$F = \beta \log \frac{\pi_{\theta}(y_p|x)}{\pi_{ref}(y_p|x)} - \beta \frac{\pi_{\theta}(y_{np}|x)}{\pi_{ref}(y_{np}|x)} \quad (2)$$

$$\mathcal{L}_{DPO}(\pi_{\theta}; \pi_{ref}) = -E_{(y_p, y_{np}) \sim \mathcal{D}} [\log \sigma(F)] \quad (3)$$

5 Experiments, Results, and Analyses

In this section, we discuss our obtained results from a single run and the effect of each module in MH-COPILOT, followed by human evaluation and analysis. Additional details are shown in Appendix E.

Performance Comparison and Ablation. We present the performance comparison of our proposed framework, MH-COPILOT, followed by an ablation study with all the contributory modules of MH-COPILOT. Table 4 presents the performance of *CSpan* module for the span prediction task. Evidently, RoBERTa-large, outperformed other encoder models on three out of four metrics. It achieves a score of 45.29%, 62.50%, and 52.52% on precision, recall, and F1, respectively. These scores observe a clear increment of +21.32%, +3.35%, and +13.77% as compared to the second-best method. Whereas, Table 5 shows the comparison of *intensity-classifier*. Here, considering the ordinality in labels, we experiment with ordinal loss and MSE loss to achieve a best F1 of 77.4%.

For the final task, generating questions, we experiment with a series of LMs. First, in order to select the base LM of MH-COPILOT, we observe zero-shot results as presented in Table 6. Evidently, Gemma-2 outperformed Llama-3, Mistral, and Phi-3 across the complete suite of metrics with a notable BERTScore (F1) of 88.58 and RL score of 42.03. We reassess their performance in a supervised-finetuning (SFT) setup and observe a mixed behavior, where Llama-3 and Mistral yield competing scores. However, we utilize Gemma-2 as our foundational model in MH-COPILOT due to its superior quantitative and qualitative performance. Finally, MH-COPILOT achieves improved scores across all evaluation metrics, with improve-

ment in the range of +2.34% to +35.74%. Last row of Table 6 reports percent improvement of MH-COPILOT against the best SFT baseline.

Ablation Study. Our ablation study assesses the contribution of two major modules in MH-COPILOT: verifier (reward model) and taxonomy. We present experimental results in Table 7. We observe that MH-COPILOT suffers a minute loss in performance when the verifier (Rew) is removed, with performance not getting impacted quantitatively. This raises questions on the impact of reward modeling in MH-COPILOT. However, qualitative analysis and human evaluation (Section 5) reveal substantial improvements in generation quality with reward modeling. On the other hand, when removing both verifier and taxonomy, we see a significant decline in the model’s performance, showing the essence of taxonomy in the whole generative framework. For instance, R1 score declines from 89.30 to 68.20, and BERTScore (F1) drops from 98.74 to 95.48. Surprisingly, Mistral+CUETAXO yields performance comparable to MH-COPILOT. Later, on qualitative analysis (Appendix G), we observed limitations in the generations of Mistral, especially pertaining to the context and LLM’s understanding of the relation between support attributes.

In general, all language models in our study see a significant jump in performance with the addition of our modules (Table 7). For instance, if we consider BERTScore (F1), then we observe improvement of +2.46 points, +3.00 points, and +2.55 points on Mistral, Phi-3, and Llama-3, respectively, on the addition of both modules.

Qualitative and Error Analysis. To further assess the quality of generations, we perform a qualitative analysis of the model’s generation. Also, we aim to justify the significance of MH-COPILOT’s reward module, *verifier*, despite observing only a small improvement in the quantitative metrics. We present a sample instance from our framework in Table 8. In this case, MH-COPILOT is able to identify that the *event* attribute is moderately present and can be further improved. On the other hand, the *effect* and *requirement* questions are direct and more aligned with the context. Evaluators attest that there exist cases, where MH-COPILOT performs better than the gold standard.

To further assess the credibility of the model’s performance and metric selection, we show additional examples and analyze them in Table 14

Models	Accuracy	Precision	Recall	F1
T5-Large (Raffel et al., 2020)	66.89	12.54	36.36	18.65
SpanBERT (Joshi et al., 2020)	69.59	22.70	33.66	27.12
RoBERTa-base	74.95	37.33	60.47	46.16
RoBERTa-large	73.00	45.29	62.50	52.52

Table 4: Span prediction results from CSpan.

Models	Accuracy		F1	
	ORD	MSE	ORD	MSE
BERT (Devlin et al., 2019)	84.00	82.30	72.00	73.10
MentalBERT (Ji et al., 2022)	85.70	83.70	73.20	68.10
MentalRoBERTa	84.80	83.20	71.20	68.30
RoBERTa	86.60	83.30	77.40	71.80

Table 5: Attribute intensity classifier results.

Models		R1	R2	RL	B1	B2	B3	B4	BERTScore			METEOR
									P	R	F1	
Mistral (Jiang et al., 2023)	Zero Shot	49.15	37.47	47.58	50.69	45.52	40.95	35.68	87.01	89.77	88.36	61.63
Phi-3 (Abdin et al., 2024)		39.22	26.98	36.29	43.80	40.04	37.48	34.46	85.28	89.47	87.32	50.09
Llama-3 (Grattafiori et al., 2024)		43.60	28.00	40.80	50.10	45.30	42.20	38.61	87.00	89.12	88.00	48.50
Gemma-2 (Riviere et al., 2024)		45.24	31.64	42.03	52.22	48.23	45.48	42.41	87.27	89.96	88.58	61.51
Mistral (Jiang et al., 2023)	SFT	<u>72.39</u>	<u>62.25</u>	<u>69.71</u>	82.25	<u>80.00</u>	<u>78.12</u>	<u>76.21</u>	<u>96.55</u>	<u>95.58</u>	<u>96.04</u>	<u>79.83</u>
Phi-3 (Abdin et al., 2024)		66.88	56.66	64.14	78.97	76.48	74.57	72.69	96.07	94.53	95.25	75.65
Llama-3 (Grattafiori et al., 2024)		71.30	61.30	68.40	<u>82.30</u>	79.70	77.82	75.91	96.30	95.30	95.80	79.00
Gemma-2 (Riviere et al., 2024)		68.20	58.04	65.58	80.39	77.87	75.95	74.00	96.10	94.94	95.48	76.98
MH-COPILOT	SFT + CUE-TAXO + Rew	89.30	84.50	88.88	93.84	92.36	91.12	89.78	98.81	98.68	98.74	93.84
$\Delta_{\text{MH-COPILOT} - \text{SFT}}(\%)$		$\uparrow 23.35$	$\uparrow 35.74$	$\uparrow 27.49$	$\uparrow 14.02$	$\uparrow 15.45$	$\uparrow 16.64$	$\uparrow 17.80$	$\uparrow 2.34$	$\uparrow 3.24$	$\uparrow 2.81$	$\uparrow 17.54$

Table 6: Performance comparison of the proposed method, MH-COPILOT, with four LLMs in two settings, including **zero-shot** and supervised-finetuning (**SFT**). Our proposed version, **MH-COPILOT**, is built upon Gemma-2, encompasses controlled taxonomy-based generations (**CUE-TAXO**), and is tuned using reward modeling (**Rew**). The last row presents our approach’s improvement (in %) with the best variant in the SFT category instead of the zero-shot category for a fair comparison. We employ standard generation metrics: ROUGE (R-1, R-2, and R-L), Meteor, BLEU (B-1, B-2, B-3, and B-4), and BERTScore (BS).

(c.f. Section I; Appendix). Although BLEU and ROUGE may reward lexical similarity (e.g., from template structure), MH-COPILOT consistently improves semantic relevance by steering the model to engage with under-addressed or missing entity types. This supports our broader claim that the MH-Copilot pipeline, guided by taxonomy-informed prompts, enhances generation quality.

Human Evaluation. Findings from Table 7 show that adding the verifier yields only a slight boost to the performance, questioning reward modeling’s impact. Although verifier modestly improves generation quality, we conduct a human evaluation to more comprehensively analyze it. The aim is to assess the generic MH-COPILOT’s generation as well as compares it with *non-verifier* variant. Our human evaluation involves scoring the MH-COPILOT’s generation based on three domain-centric parameters and four linguistic parameters. Domain-centric evaluation metrics are defined as

Models	R1	R2	RL	B1	B2	B3	B4	BS-F1	MET
MH-COPILOT	89.30	84.50	88.88	93.84	92.36	91.12	89.78	98.74	93.84
–Rew	88.32	82.97	87.67	93.00	91.25	89.83	88.32	98.52	93.33
–Rew–CUE-TAXO	68.20	58.04	65.58	80.39	77.87	75.95	74.00	95.48	76.98
Mistral (SFT)	72.39	62.25	69.71	82.25	80.00	78.12	76.21	96.04	79.83
+CUE-TAXO	88.98	83.77	88.31	93.49	91.77	90.37	88.88	98.68	93.57
+CUE-TAXO+Rew	88.14	82.24	87.42	92.49	90.50	89.00	87.52	98.50	92.60
Phi-3 (SFT)	66.88	56.66	64.14	78.97	76.48	74.57	72.69	95.25	75.65
+CUE-TAXO	86.80	80.94	85.98	92.04	90.10	88.69	87.10	98.33	92.16
+CUE-TAXO+Rew	86.56	80.60	85.66	92.00	90.10	88.53	87.00	98.25	91.88
Llama-3 (SFT)	71.30	61.30	68.40	82.30	79.70	77.82	75.91	95.80	79.00
+CUE-TAXO	87.46	81.96	86.83	92.49	90.70	89.26	87.71	98.28	94.70
+CUE-TAXO+Rew	88.28	82.93	87.70	93.17	91.41	89.98	88.48	98.35	93.23

Table 7: **Ablation Study.** We demonstrate the effectiveness of each model component: taxonomy (**CUE-TAXO**), and verifier for reward modeling (**Rew**).

follows:

- D1: Contextual relevance** of the generated questions with regard to the given input post;
- D2: Empathy** shows how well the generated questions avoid triggering the support seeker;
- D3: Value addition** justifies how much is the generated question’s response likely to add more value to the original post.

On the other hand, linguistic evaluation metrics are defined as follows:

- L1: Relevance** measures the selection of relevant content considering the reference utterance.
- L2: Coherence** examines the structure and organization of the generated responses;
- L3: Fluency** assesses the linguistic quality of the generated responses;
- L4: Consistency** evaluates factual alignment with the source utterance.

We curated a sample of 40 random instances, and asked evaluators to rate linguistic metrics (L-1,2,3,4) on a scale of 1 (worst) to 5 (best) and domain-centric metrics (D-1,2,3) on a scale of 1 to 3. Evaluators are from mixed professions, aged between 21-35, and have been previous users of targeted subreddit channels. We repeat the same process for *non-verifier* setup to add justification for the relevance of reward modeling from a qualitative aspect. As shown in Table 9, the Domain section presents domain-centric results, whereas the Linguistic section presents linguistic results. Evidently, MH-COPILOT excels across all domain-centric and

Original Post: *So I have been taking adderall 10mg for a little over a week and I've been testing it. Today i decided to take adderall and coffee at the same time. I think this was a bad idea because i felt extremely jittery and anxious, but the worst part was i came down from them at the same time. I am not fully sure of my personal side effects of adderall yet but coming down at the same time from both i had this weird feeling of nausea, weak, bad headache, tired, extremely zoned out, and almost like a not knowing what's going on kinda feeling, also i'm just getting over the flu. Not sure why this happened but it sucked. Should I just avoid caffeine for now on, I can do that cause the adderall is doing way more than the caffeine was. The reason i'm asking is cause I couldn't find anyone having this problem. So if you guys got any tips that would be helpful, like if its the flu, or the adderall, or the caffeine, or if anyone else has experienced this.*

⇒ **Annotated:** ⟨ Event Level: Moderate ⟩ ⟨ Effect Level: Well-described ⟩ ⟨ Requirement Level: Well-described ⟩
 ⟨ **Event:** Can you elaborate more on why you are taking adderall? ⟩ ; ⟨ **Effect:** n/a ⟩ ; ⟨ **Requirement:** n/a ⟩

⇒ **MH-COPILOT:** ⟨ Event Level: Moderate ⟩ ⟨ Effect Level: Moderate ⟩ ⟨ Requirement Level: Well-described ⟩
 ⟨ **Event:** Can you elaborate more on the Adderall? ⟩ ; ⟨ **Effect:** Can you elaborate more on how you feel after taking Adderall? ⟩ ; ⟨ **Req:** n/a ⟩

Table 8: An error analysis of MH-COPILOT. *Event*, *Effect*, and *Requirement* spans are illustrated in magenta, blue, and cyan, respectively. [Best viewed in color]. More examples in Appendix F.

	Domain			Linguistic			
	D1	D2	D3	L1	L2	L3	L4
w/o Verifier	3.27	1.82	2.19	3.46	3.70	3.82	3.80
w/ Verifier	3.43	2.27	3.31	3.62	3.84	4.02	3.89

Table 9: Human evaluation on the responses generated from MH-COPILOT when compared to MH-COPILOT without verifier. The performance of MH-COPILOT across all metrics is better than without verifier module.

linguistic metrics. On linguistic aspects, generations are fluent (4.02) as well as coherent (3.84), as fluency and coherency receive maximum human feedback rating after adding the *verifier* module. This further indicates that the *verifier* module outputs outperform the *non-verifier* setup, showing qualitative superiority. Additionally, annotators further state that the generations from MH-COPILOT, in some cases, surpass the gold standard.

6 Conclusion

To address the persistent issue of low response rates on support seeker’s posts in OMHCs, we proposed MH-COPILOT, a reinforcement-learning-based framework designed to assess and enhance the clarity of support-seeking posts. Our contributions include REDDME, a novel dataset comprising 4,760 posts annotated with key support attributes event, effect, and requirement at both span and intensity levels. Additionally, we propose a dedicated taxonomy, CUETAXO, for controlled prompting and generation. Next, we propose MH-COPILOT that employs CUETAXO to generate contextually relevant and tailored questions, prompting users to provide more complete and actionable information to improve their posts in order to elicit more responses from support-providers. Our extensive benchmarking against notable language models demonstrated consistent improvements across 11 metrics. We concluded our work with an exhaus-

tive analysis supplemented with human evaluation.

7 Ethical Considerations

Our work acknowledges that support-seekers often face disparities in engagement, whereas those who articulate their issues clearly tend to receive responses, while others remain unnoticed. It is worth noting that our system is not intended to ‘diagnose’ or ‘suggest’ remedies for mental health issues. Its sole purpose is to help users articulate their thoughts more clearly, enabling a better peer support ecosystem. Our system is designed to be non-intrusive and supportive, which gradually encourages users in crisis to seek peer help. We provide additional considerations in Appendix H.

8 Limitations and Future Scope

While MH-COPILOT demonstrates significant improvements in enhancing support-seeking posts within OMHCs, it has limitations, including reliance on the specific dataset and potential biases inherent in language models. Future work could explore expanding the dataset to include diverse online platforms and refining the model to better address nuances in user expression. Furthermore, MH-COPILOT is inspired by therapeutic principles and aligns with prior work on explainable and controllable generation in sensitive domains like mental health. The modularity also allows for extensibility to other domains. Additionally, integrating real-time feedback mechanisms and user-centered evaluations could further enhance the model’s applicability and effectiveness in diverse mental health contexts.

9 Acknowledgement

The authors acknowledge the support of the Infosys Foundation through CAI at IIIT-Delhi.

References

- Marah Abidin, Sam Ade Jacobs, and Ammar Ahmad Awan. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *ArXiv*, abs/2404.14219.
- Protay Kumar Adhikary, Aseem Srivastava, Shivani Kumar, Salam Michael Singh, Puneet Manuja, Jini K Gopinath, Vijay Krishnan, Swati Kedia Gupta, Koushik Sinha Deb, and Tanmoy Chakraborty. 2024. [Exploring the efficacy of large language models in summarizing mental health counseling sessions: Benchmark study](#). *JMIR Ment Health*, 11:e57306.
- 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.
- C. Benjet, E. Bromet, E. G. Karam, R. C. Kessler, K. A. McLaughlin, A. M. Ruscio, V. Shahly, D. J. Stein, M. Petukhova, E. Hill, and et al. 2016. [The epidemiology of traumatic event exposure worldwide: results from the world mental health survey consortium](#). *Psychological Medicine*, 46(2):327–343.
- Ziyi Chen, Ren Yang, Sunyang Fu, Nansu Zong, Hongfang Liu, and Ming Huang. 2023. [Detecting reddit users with depression using a hybrid neural network sbert-cnn](#). In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*. IEEE.
- Kendra Cherry and David Susman. 2023. [What are the big 5 personality traits?](#) Verywell Mind.
- Nikki Ciletti. 2025. [Steps in therapy: What to expect during the therapeutic process](#). BetterHelp.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Chandra Kiran Evuru, Sreyan Ghosh, Sonal Kumar, Rameshwaran S, Utkarsh Tyagi, and Dinesh Manocha. 2024. [CoDa: Constrained generation based data augmentation for low-resource NLP](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3754–3769, Mexico City, Mexico. Association for Computational Linguistics.
- Muskan Garg, Manas Gaur, Raxit Goswami, and Sunghwan Sohn. 2023. [Lost: A mental health dataset of low self-esteem in reddit posts](#). In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3854–3859.
- Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. ["let me tell you about your mental health!": Contextualized classification of reddit posts to dsm-5 for web-based intervention](#). In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18*, page 753–762, New York, NY, USA. Association for Computing Machinery.
- Naline Geurtzen, Ger P.J. Keijsers, Johan C. Karremans, Bea G. Tiemens, and Giel J.M. Hutschemaekers. 2020. [Patients' perceived lack of goal clarity in psychological treatments: Scale development and negative correlates](#). *Clinical Psychology & Psychotherapy*, 27(6):915–924.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle. 2024. [The llama 3 herd of models](#). *ArXiv*, abs/2407.21783.
- Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordani, and Rishabh Agarwal. 2024. [V-Star: Training verifiers for self-taught reasoners](#). In *First Conference on Language Modeling*.
- Yusif Ibrahimov, Tarique Anwar, and Tommy Yuan. 2024. [Explainable ai for mental disorder detection via social media: A survey and outlook](#). *ArXiv*, abs/2406.05984.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. [MentalBERT: Publicly available pretrained language models for mental healthcare](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.
- Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, and Diego de Las Casas. 2023. [Mistral 7b](#). *ArXiv*, abs/2310.06825.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. [SpanBERT: Improving pre-training by representing and predicting spans](#). *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Seoyun Kim, Junyeop Cha, Dongjae Kim, and Eunil Park. 2023. [Understanding mental health issues in different subdomains of social networking services: Computational analysis of text-based reddit posts](#). *J Med Internet Res*, 25:e49074.
- James Kirkbride, Deidre Anglin, Ian Colman, Jen Dykshoorn, Peter Jones, Praveetha Patalay, Alexandra Pitman, Emma Soneson, Thomas Steare, Talen Wright,

- and Sian Griffiths. 2024. [The social determinants of mental health and disorder: evidence, prevention and recommendations](#). *World Psychiatry*, 23:58–90.
- Shanu Kumar, Akhila Yesantarao Venkata, Shubhanshu Khandelwal, Bishal Santra, Parag Agrawal, and Manish Gupta. 2024. [Sculpt: Systematic tuning of long prompts](#). *arXiv preprint arXiv:2410.20788*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *ArXiv*, abs/1907.11692.
- Usha Lokala, Aseem Srivastava, Triyasha Ghosh Dastidar, Tanmoy Chakraborty, Md Shad Akhtar, Maryam Panahiazar, and Amit Sheth. 2022. [A computational approach to understand mental health from reddit: Knowledge-aware multitask learning framework](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):640–650.
- J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, and S. J. Bartels. 2016. [The future of mental health care: peer-to-peer support and social media](#). *Epidemiology and Psychiatric Sciences*, 25(2):113–122.
- Mark Olfson. 2016. [Building the mental health workforce capacity needed to treat adults with serious mental illnesses](#). *Health Aff (Millwood)*.
- Flavio Petruzzellis, Alberto Testolin, and Alessandro Sperduti. 2024. [Benchmarking GPT-4 on algorithmic problems: A systematic evaluation of prompting strategies](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2161–2177, Torino, Italia. ELRA and ICCL.
- Yada Pruksachatkun, Sachin R. Pendse, and Amit Sharma. 2019. [Moments of change: Analyzing peer-based cognitive support in online mental health forums](#). In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–13, New York, NY, USA. Association for Computing Machinery.
- 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](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Sunny Rai, Khushi Shelat, Devansh R Jain, Kishen Sivabalan, Young Min Cho, Maitreyi Redkar, Samindara Sawant, and Lyle H. Ungar Sharath Chandra Guntuku. 2024. [Cross-cultural differences in mental health expressions on social media](#). *ArXiv*, abs/2402.11477.
- Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, and Surya Bhatiraju. 2024. [Gemma 2: Improving open language models at a practical size](#). *ArXiv*, abs/2408.00118.
- Brian Romanowski, Asma Ben Abacha, and Yadan Fan. 2023. [Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches](#). *Journal of the American Medical Informatics Association*, 30(8):1448–1455.
- Chandni Saxena, Muskan Garg, and Gunjan Ansari. 2022. [Explainable causal analysis of mental health on social media data](#). In *Neural Information Processing - 29th International Conference, ICONIP 2022, Virtual Event, November 22-26, 2022, Proceedings, Part II*, volume 13624 of *Lecture Notes in Computer Science*, pages 172–183. Springer.
- Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020a. [Engagement patterns of peer-to-peer interactions on mental health platforms](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):614–625.
- 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*, WWW ’21, page 194–205, New York, NY, USA. Association for Computing Machinery.
- Ashish Sharma, Inna Wanyin Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2022. [Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support](#). *Nature Machine Intelligence*, 5:46–57.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020b. [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, Online. Association for Computational Linguistics.
- Aseem Srivastava, Tanya Gupta, Alison Cerezo, Sarah Peregrine (Grin) Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2025a. [Critical behavioral traits foster peer engagement in online mental health communities](#). *PLOS ONE*, 20(1):1–17.
- Aseem Srivastava, Gauri Naik, Alison Cerezo, Tanmoy Chakraborty, and Md. Shad Akhtar. 2025b. [Sentiment-guided commonsense-aware response generation for mental health counseling](#). *Preprint*, arXiv:2501.03088.
- Aseem Srivastava, Ishan Pandey, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. [Response-act guided reinforced dialogue generation for mental health counseling](#). In *Proceedings of the ACM Web Conference*

2023, WWW '23, page 1118–1129, New York, NY, USA. Association for Computing Machinery.

Victoria Stewart, Sara S McMillan, Jie Hu, Jack C Collins, Sarira El-Den, Claire L O'Reilly, and Amanda J Wheeler. 2024. [Are smart goals fit-for-purpose? goal planning with mental health service-users in australian community pharmacies.](#) *International Journal for Quality in Health Care*, 36(1):mzae009.

Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. [Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills.](#) *J Med Internet Res*, 21(7):e12529.

Sayali C. Tiwari and Swati R. Deshpande. 2020. [A study to assess the effect of stressful life events on psychological distress levels of participants living in an urban area.](#) *Journal of Family Medicine and Primary Care*, 9(6):2730–2735.

Leandro Von Werra, Lewis Tunstall, Abhishek Thakur, Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, and Helen Ngo. 2022. [Evaluate & evaluation on the hub: Better best practices for data and model measurements.](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 128–136, Abu Dhabi, UAE. Association for Computational Linguistics.

Joseph Weizenbaum. 1966. [Eliza—a computer program for the study of natural language communication between man and machine.](#) *Commun. ACM*, 9(1):36–45.

Marsha White and Steve M. Dorman. 2001. [Receiving social support online: implications for health education.](#) *Health Education Research*, 16(6):693–707.

Diya 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*, CHI '19, page 1–15, New York, NY, USA. Association for Computing Machinery.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. [Star: self-taught reasoner bootstrapping reasoning with reasoning.](#) In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. [Finding your voice: The linguistic development of mental health counselors.](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 936–947, Florence, Italy. Association for Computational Linguistics.

Overview of Appendices

- Appendix A: Additional Annotation Rationale.
- Appendix B: Annotator Guidelines.
- Appendix C: Dataset Analysis.
- Appendix D: Prompts.
- Appendix E: Training Details.
- Appendix F: Additional Examples.
- Appendix G: Comparison of Mistral and MH-COPILOT.
- Appendix H: Additional Ethical Considerations.
- Appendix I: Additional Analysis.

A Additional Annotation Rationale

In the annotation process, we are required to identify the precise support attributes that potentially appeal to support providers to respond. Considering our goal of proposing a solution that does not target just one metric and its improvement but a more holistic improvement in support-seeker experiences, we realize that generating high-quality suggestions for the users to improve their posts would require the rationales governing the annotator's judgment for the presence or absence of the critical elements. Moreover, at times, support seekers touch upon the subject without providing adequate insights, prompting the support providers to respond in an abstract way. Therefore, we also mark the intensity/level of their presence in the post. Finally, to provide proactive support, we annotate reference questions that would encourage the user to write an appropriate post considering their current attribute-level response.

B Annotator Guidelines

In this annotation task, identify and label spans within Reddit posts that express emotional distress, mental health symptoms, coping mechanisms, or recovery efforts. Focus on subjective, psychologically significant content such as feelings of sadness, anxiety, hopelessness, self-perception, mentions of therapy or medication, or descriptions of distressing experiences. For each span, assign an intensity score from 0 to 2. Avoid annotating spans that are purely narrative, background, or unrelated to mental health. If a span includes a long sentence

#	Subreddit	Description
1	r/Anxiety	A subreddit for support and discussion around anxiety disorders.
2	r/ptsd	A community for individuals dealing with post-traumatic stress disorder.
3	r/addiction	Support and resources for individuals battling addiction.
4	r/ADHD	A subreddit focused on attention-deficit/hyperactivity disorder discussions.
5	r/alcoholicsanonymous	A forum for members of Alcoholics Anonymous and those interested.
6	r/Anger	A community to discuss and manage anger issues.
7	r/BPD	A space for individuals with borderline personality disorder.
8	r/depression	Support and discussion for those dealing with depression.
9	r/domesticviolence	A community for survivors of domestic violence seeking support.
10	r/getting_over_it	A space for advice on overcoming challenges and struggles.
11	r/mentalillness	Discussions and support around various mental illnesses.
12	r/OpiatesRecovery	A subreddit for recovery from opiate addiction.
13	r/rapecounseling	Support and counseling for survivors of sexual violence.
14	r/sad	A community for sharing and coping with feelings of sadness.
15	r/selfharm	A supportive space for those dealing with self-harm.
16	r/selfhelp	Resources and discussions on self-improvement and help.
17	r/socialanxiety	A subreddit for individuals with social anxiety disorder.

Table 10: Description of the various subreddits used in the BeCOPE (Srivastava et al., 2025a) and also the REDDME dataset.

expressing different attributes (event, effect or requirement), break the sentence using suitable punctuation and annotate them separately. Be consistent, base your judgment on the text itself, and flag any uncertain or potentially serious content for review if required. Include the title for annotation if the title includes a significant context.

C Dataset Analysis

Table 10 presents a comprehensive list of the subreddits included in the REDDME dataset. These subreddits were carefully selected to provide a diverse and representative sample of online discourse across a wide range of topics and communities. The selection process aimed to capture variations in language use, sentiment, community norms, and topical focus, all of which are crucial for robust linguistic and social analysis.

From the source BeCOPE dataset, we utilize interactive posts from 17 subreddits, which constitutes a total of 4760 posts. The final dataset, REDDME, contains post title, post body, annotated post body, event-intensity, effect-intensity, requirement-intensity, event-question, effect-question, and requirement-question. The average post length is 179.62 words, and we observe a total of 2125 posts in which the event is absent or moderately present, whereas 2781 posts for effect and 2976 posts for requirement, where we observe it to be absent or moderately present. Thus, a large percentage of the posts in the dataset either have a complete absence or only a moderate presence of the attributes. This indicates that the majority of support seekers are unable to properly express their problems. Our annotators’ marked rationale’s average event-span-length is 65.70 words, 26.48

words for effect-span-length, and 19.31 words for requirement-span-length. This marks that the support seekers have mainly written about the triggering incidents but have provided very little information about the supporting effects of that incident and what they want as help from the community. The post-body length is observed to follow an increasing trend as the intensity levels increase for each cue.

D Prompts

Table 11 describes the user and system prompts used for the Language Models for generating the guiding questions. The system prompt briefly describes the role of a Online Mental Health Platforms, support providers and support seekers.

The user prompt consists of the reddit post with event, effect and requirement spans as well as their respective intensity levels.

E Training Details

We fine-tune four LMs: LLaMA 3-instruct (8B), Phi-3-mini-4k-instruct (3.8B), Mistral-instruct-v0.1 (7B), and Gemma-2-instruct (2B) using the QLoRA (Dettmers et al., 2024) method (quantization=8 bit, r=16, alpha=32, dropout=0.05). The context length was set to 1024 tokens, optimized for a balance between the model’s limit and computational efficiency. Training is carried out on 1x NVIDIA A100 GPUs (80GB) and 1x RTX A6000 GPUs (50GB), ensuring sufficient capacity for gradient accumulation and model checkpoints. We use a learning rate of 2e-5 for SFT to achieve stable convergence with AdamW optimizer. The training was carried out for 2 epochs after which the models started overfitting. We employed the

System Prompt

A support seeker on a peer-to-peer (P2P) Online Mental Health Platform (OMHP) is an individual who utilizes digital services to seek assistance/help for managing and improving their mental health, typically through interactions with peer groups or self-help resources.

The parameters are defined as follows:

Event: This parameter encapsulates the specific situation, activity, or event that is the focal point of the support seeker’s concern. The explicit detailing of such events provides a contextual background essential for empathetic understanding.

Effect: This aspect targets the impact or consequences of the identified event on the support seeker. By elucidating the effect, the post conveys the emotional or practical repercussions of the event, thereby inviting more targeted and empathetic responses.

Requirement: This parameter is critical in directing the nature of the assistance sought. It ranges from emotional and informational support to instrumental aid, thereby guiding the potential response trajectory.

In the posts on OMHP, these parameters can have intensity ranging from 0 to 2, where 0 means absent, 1 means present but needs clarification, and 2 being well-described based on the presence of these parameters in the post.

Consider the following post by a support seeker on a OMHP, in which the spans of text representing *event*, *effect*, and *requirement* have been marked. Also, the intensity levels for each of the parameters in the post have been provided along with the post. The post is context of the victim. The `<es>` and `<ee>` tags encapsulate the spans for the *event* parameter, `<efs>` and `<efe>` tags encapsulate the spans for the *effect* parameter, and `<rs>` and `<re>` tags encapsulate the spans for the *requirement* parameter.

User Prompt

Post: <Post Body>

Event scale: **Effect scale:** **Requirement scale:**

Schema: {event_question: , effect_question: , requirement_question: }

Generate 3 questions following the schema provided below the post, for helping the support giver to understand more about the victim. Strictly follow the question format of schema. Give only the json output as specified in the schema and no explanation needed.

Table 11: The prompts used by language models to generate the guiding questions. The system prompt is a description of the task and description of all the attributes. The user prompt contains the Reddit post, intensity scale and schema for the generation format.

BLEU, ROUGE and METEOR metrics with default parameters from the Hugging Face² Evaluate library(Von Werra et al., 2022) for our evaluations. All experiments were run under Python 3.8.19, with PyTorch 2.4.1 and Transformers 4.45.0³.

F Additional Examples

Table 12 presents representative outputs from the full MH-COPILOT pipeline. For each support attribute, we display the extracted text spans, their predicted intensity scores, and an associated guiding question designed to probe that attribute.

G Comparison of Mistral and MH-COPILOT

Table 13 shows a comparative study between the generations of Mistral and MH-COPILOT. It clearly shows that RLHF improved the quality of the generated guiding questions. In case #1, MH-COPILOT is able to identify that the event attribute is moderately present and can be further improved. On the other hand, the effect and requirement questions are direct and more aligned with the context. Similarly, case #2 shows that the support seeker had already described much about the effect attribute on the topic related to ‘adderall’ but MH-COPILOT

has incorrectly classified its intensity and generated a question under that category. In this case, the fluency of the event question has also suffered. This proves that the model’s performance remains intact till the pipeline’s other modules remain intact, as it’s evident that errors at the early phase get escalated to subsequent steps.

H Additional Ethical Considerations

Our proposed dataset is an extension of the publicly available BeCOPE dataset, which is anonymized at the source with all personally identifiable information (PII) removed by its original creators. We ensured that this standard of anonymization was strictly maintained in our extension. All human contributors involved in annotation and evaluation were adults who participated voluntarily with informed consent, having been clearly briefed beforehand about the nature of the task and the potentially sensitive content they might encounter. Resources were provided to help them understand the problem, and participation remained optional at every stage. The resulting REDDME dataset will be released publicly for research purposes under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license, which allows sharing and adaptation for non-commercial use with proper attribution. To further safeguard

²<https://huggingface.co/>

³<https://huggingface.co/docs/transformers>

#	Summaries
1.	OP: <i>So I have been taking adderall 10mg for a little over a week and I've been testing it. Today i decided to take adderall and coffee at the same time. I think this was a bad idea because i felt extremely jittery and anxious, but the worst part was i came down from them at the same time. I am not fully sure of my personal side effects of adderall yet but coming down at the same time from both I had this weird feeling of nausea, weak, bad headache, tired, extremely zoned out, and almost like a not knowing what's going on kinda feeling, also i'm just getting over the flu. Not sure why this happened but it sucked. Should I just avoid caffeine for now on, I can do that cause the adderall is doing way more than the caffeine was. The reason i'm asking is cause I couldn't find anyone having this problem. So if you guys got any tips that would be helpful, like if its the flu, or the adderall, or the caffeine, or if anyone else has experienced this.</i> ⇒ Annotated: { Event Level: Moderate } { Effect Level: Well-described } { Requirement Level: Well-described } { Event: Can you elaborate more on why you are taking adderall? } ; { Effect: n/a } ; { Requirement: n/a } ⇒ MH-COPILOT: Event Level: Moderate Effect Level: Moderate Requirement Level: Well-described { Event: Can you elaborate more on the Adderall? } ; { Effect: Can you elaborate more on how you feel after taking Adderall? } ; { Req: n/a }
2.	OP: <i>Long story short - Am I completely bonkers here? I get that high energy exercise is good for improving dopamine and norepinephrine recommended, which is for several reasons but one big one is I get winded climbing short staircases. I'm 250lbs and haven't even walked for exercise in 6 months so I feel it's a terrible idea to jump directly into HIIT.</i> <i>Am I completely bonkers here? I get that high energy exercise is good for improving dopamine and norepinephrine levels and have had good results from interval training in the past, but I also don't want to injure myself or trigger a heart attack. I have promised to do Yoga for a while, starting slow, and then looking into HIIT later but he called me lazy and said I'm hiding behind my diagnosis to avoid changing bad habits. :(He's not usually so mean, but he's been very pushy about this. I don't know what to do.</i> ⇒ Annotated: { Event Level: Well-described } { Effect Level: Absent } { Requirement Level: Moderate } { Event: n/a } ; { Effect: How did doing yoga make you feel? } ; { Req: Can you elaborate more on what can help you change your bad habits? } ⇒ MH-COPILOT: Event Level: Moderate Effect Level: Moderate Requirement Level: Absent { Event: Can you elaborate more on your diagnosis? } ; { Event: Can you elaborate more on your diagnosis? } ; { Event: Can you elaborate more on your diagnosis? }
3.	OP: <i>Hey everyone, I'd really appreciate some assistance with this one so please stick with me. A bit of back ground; I'm a 25 year old female, I've been diagnosed with BPD before someone else diagnosed me with Bipolar type 2. I've always have massive issues with anxiety and recently OCD, i also had some self harming tendencies. But one thing really bugs me. Since I was a young teen I've had these extremely in depth fantasies about being someone else, not that odd I know, but it's to a crazy extent. I've never brought it up because honestly I'm embarrassed. Always about being this one person I've created in my head, sometimes I fantasize about being them at different stages of their life. But always the same person. I have created all these imaginary friends and relationship that this persons encounters. My brain acts out full conversations and everything...</i> <i>I thought maybe this had begun as a coping mechanism after being so alone for so long when I was younger. (Btw were talking about living in these fantasies for multiple hours a day, while I work, while I drive, sometimes I can't even concentrate on a movie because of them) it's all contained in my head and no one knows it's happening.</i> <i>But sometimes I'll catch myself making a facial expression/ sound or movement that's related to whats happening in my head not real life...</i> <i>I have friends now and life's okay, so I don't know why I still do this, I'm frightened there may be something wrong with my brain that's laying dormant...</i> <i>So does anyone have a name for what this may be?</i> <i>If you read all of this, thank you so much, really.</i> ⇒ Annotated: { Event Level: Well-described } { Effect Level: Moderate } { Requirement Level: Moderate } { Event: n/a } ; { Effect: Can you elaborate more on how the fantasies make you feel? } ; { Requirement: Can you elaborate more on what would help you overcome these delusions? } ⇒ MH-COPILOT: Event Level: Moderate Effect Level: Moderate Requirement Level: Absent { Event: Can you elaborate more on your fantasies? } ; { Effect: Can you elaborate more on how you feel about your fantasies? } ; { Req: What do you need help with now that you are having fantasies about being someone }

Table 12: [Best viewed in color] A comparative analysis of positive and error cases of MH-COPILOT. Case #1 and Case #3 demonstrate the ability to capture insightful nuances, whereas Case #2 demonstrates erroneous generation escalated from misclassification. *Event* is illustrated in magenta, *effect* spans in blue, and *requirement* spans in cyan. Additionally *Correct attribute intensity* are shown in green and *incorrect attribute intensity* in red.

ethical use, researchers accessing the dataset will be required to agree not to use it for commercial purposes. The BeCOPE dataset, from which our work extends, is already publicly available under similar guidelines.

I Additional Analysis

To further assess the credibility of the selection of metrics and the model's performance, we perform additional qualitative analysis. Table 14 shows supporting examples for this analysis.

- In the first example from our dataset, the Effect and Requirement entities are already well described (Level 2), whereas the Event entity is

only minimally present (Level 1). Without Cue-Taxo, the model generates a follow-up targeting the Effect entity, already well-formed, thus missing an opportunity to probe the underdeveloped Event. With CueTaxo, the follow-up instead targets the Event, resulting in a more contextually useful and balanced question. This shift corresponds to an increase in both BLEU and ROUGE scores, aligning with improved semantic targeting.

- In the second example, the Requirement entity is completely missing from the post. Without Cue-Taxo, the model attempts to generate a follow-up targeting only this missing entity, while ignoring

#	Reddit Posts
1.	I keep having positive dreams about the person that raped me over a year ago. he has verbally abused me in public for the past year and made my life a misery. I am terrified of him. why do i keep having dreams about him where he's being nice to me and apologising for what he did, in my dreams he is a nice person but in real life he's not. please help me understand why this is happening. [Event: 2; Effect: 1; Requirement: 2] Mistral+CUETAXO: Effect question: Can you elaborate more on how you feel about the dreams? MH-COPILOT: Effect question: Can you elaborate more on how the dreams make you feel ?
2.	I loved the Psych Ward I felt so safe and surrounded by nice non-judgemental people but of course my anxiety ruined it and it just became torture I felt on the verge of throwing up 24/7 and I barely left my room and couldn't eat and every time I went for meds I shook like I had Parkinsons because my dumb brain can't handle people focusing on me They were so nice man they even covered the mirrors for my BDD I wanna go back so bad but I'll feel bad there because I feel bad everywhere and I just wanna die so bad. It's a loss loss I can't win this battle [Event: 1; Effect: 2; Requirement: 0] Mistral+CUETAXO: Event question: Can you elaborate more on what caused your anxiety? Requirement question: What do you need help with now that you feel bad everywhere? MH-COPILOT: Event question: Can you elaborate more on why you were in the psych ward? Requirement question: What do you need help with now that you feel you were tortured in the psych ward?
3.	So I have been taking adderall 10mg for a little over a week and I've been testing it.Today i decided to take adderall and coffee at the same time.I think this was a bad idea because i felt extremely jittery and anxious, but the worst part was i came down from them at the same time.I am not fully sure of my personal side effects of adderall yet but coming down at the same time from both I had this weird feeling of nausea, weak, bad headache, tired, extremely zoned out, and almost like a not knowing what's going on kinda feeling, also im just getting over the flu.Not sure why this happened but it sucked.Should I just avoid caffeine for now on, I can do that cause the adderall is doing way more than the caffeine was. The reason im asking is cause I couldn't find anyone having this problem. So if you guys got any tips that would be helpful, like if its the flu, or the adderall, or the caffeine, or if anyone else has experienced this. [Event: 1; Effect: 2; Requirement: 2] Mistral+CUETAXO: Event question: Can you elaborate more on why you took adderall and coffee at the same time ? MH-COPILOT: Event question: Can you elaborate more on why you are taking adderall?

Table 13: A comparative analysis of guiding questions generated by Mistral + CUETAXO and MH-COPILOT.

the Event (present at Level 1). CueTaxo instead guides the model to generate follow-ups for both missing and underdeveloped entities, better capturing the post's intent and producing a more complete response, again reflected in improved scores.

- In the third example, the question generated without CueTaxo is appropriate but fails to include any question related to the Effect entity. The CueTaxo-guided output includes this additional, contextually relevant query. Even though both versions may receive high BLEU/ROUGE due to templated phrasing, the CueTaxo version provides greater utility and completeness, contributing to a genuine gain in evaluation metrics.

#	Reddit Posts
1.	My partner can get very angry at times and it makes me feel uncomfortable. We now have a baby and I do not want him to be that way around her but I don't know the best way to talk to him about this. What should I do if he ever is angry around her? I want him to know that it is absolutely unacceptable if he ever is like that. Any help would be really appreciated. Without CUE TAXO: <i>Event Question:</i> <i>Effect Question:</i> How did your partner's anger make you feel? <i>Requirement Question:</i> With CUE TAXO <i>Event Question:</i> Can you elaborate more on your partner's anger? <i>Effect Question:</i> <i>Requirement Question:</i>
2.	I loved the Psych Ward. I felt so safe and surrounded by nice non-judgemental people, but of course, my anxiety ruined it and it just became torture. I felt on the verge of throwing up 24/7 and I barely left my room and couldn't eat. Every time I went for meds I shook like I had Parkinson's because my brain can't handle people focusing on me. They were so nice—they even covered the mirrors for my BDD. I wanna go back so bad but I'll feel bad there, because I feel bad everywhere. I just wanna die so bad. It's a loss-loss. I can't win this battle. Without CUE TAXO: <i>Event Question:</i> <i>Effect Question:</i> <i>Requirement Question:</i> What do you need help with now that you feel like you are losing control of your mental health? With CUE TAXO: <i>Event Question:</i> Can you elaborate more on why you were in the psych ward? <i>Effect Question:</i> <i>Requirement Question:</i> What do you need help with now that you feel you were tortured in the psych ward?
3.	I'll start off by saying some days I feel somewhat normal. But mainly I isolate myself from people and find other people annoying and sensitive... I also tend to (and like to) take things quite literally, perhaps due to the structure of programming. My apologies if this is the wrong sub for this, and my behavior described is actually quite common. Without CUE TAXO: <i>Event Question:</i> <i>Effect Question:</i> <i>Requirement Question:</i> What do you need help with now that you feel isolated from people? With CUE TAXO: <i>Event Question:</i> <i>Effect Question:</i> Can you elaborate more on how you feel about socializing? <i>Requirement Question:</i> What do you need help with now that you feel isolated from people?

Table 14: A comparative analysis of guiding questions generated without CUE TAXO and with CUE TAXO with Gemma.