# Mental Health Signal Detector: Intent, Concern Classification, and Suggestion via RAG in Online Support Groups

## Project Final Report

**Gaurav Zanpure, Shreyansh Kabra, Pragya Dhawan, Suyash Roy, Vanshika Wadhwa, Punith Basavaraj**
University of Southern California
`{gzanpure, kabras, pragyadh, suyashro, wadhwav, punithba}@usc.edu`

## Abstract

The exponential growth of user-generated content in online support communities has generated an unprecedented archive of real-time mental health data. Platforms such as Reddit host vast communities where individuals discuss depression, anxiety, and suicidal ideation with a level of candor rarely observed in clinical settings. However, the sheer volume of data overwhelms human moderation capacities, necessitating the development of scalable, ethically robust automated triage systems. This paper presents the Mental Health Signal Detector, a Natural Language Processing (NLP) framework designed to classify Reddit posts by specific Intent (e.g., Critical Risk, Seeking Help, Maladaptive Coping) and Concern Level (Low, Medium, High).

To address the scarcity of labeled data, we curated a dataset of approximately 6,000 posts, establishing a "gold standard" subset through manual annotation. We implemented a hybrid labeling pipeline that bootstraps large-scale labels by combining high-precision Python regex rules with zero-shot classification (BART-large-MNLI), refined via human-in-the-loop validation. Benchmarking three supervised models—MiniLM, DistilBERT, and DistilRoBERTa (with LoRA)—yielded macro-F1 scores of 0.70 for intent detection and 0.78 for concern assessment. Furthermore, we extended the system with a Retrieval-Augmented Generation (RAG) module to assist moderators. By leveraging SentenceTransformer embeddings and FAISS indexing to retrieve semantically relevant, safe peer-support examples, the system conditions a Flan-T5 model to generate empathetic, context-aware responses. This architecture ensures grounding in real discourse, minimizes hallucination, and provides moderators with actionable, safety-compliant guidance.

## 1 Motivation

Online support communities, specifically subreddits such as `r/depression`, `r/anxiety`, and `r/suicidewatch`, have emerged as vital, accessible avenues for individuals to vocalize emotional struggles and seek peer validation. However, the exponential influx of user-generated content in these spaces severely outpaces the bandwidth of human moderators. This discrepancy creates a critical bottleneck where high-risk posts—particularly those exhibiting immediate suicidal ideation—may remain unaddressed for prolonged periods.

To bridge this gap, there is an urgent need for automated systems that act not merely as filters, but as intelligent triage assistants. Our objective is to design an interpretable and ethically robust NLP framework capable of parsing the semantic complexity of mental health discourse. Specifically, the system must distinguish between nuanced intents (e.g., Seeking Help vs. Venting) and quantify the concern level to prioritize urgent cases. Furthermore, to mitigate the cognitive load on human moderators, the system must go beyond classification to provide actionable support. By integrating Retrieval-Augmented Generation (RAG), we aim to synthesize empathetic, context-aware response candidates that are grounded in safety guidelines, thereby ensuring psychological safety while maintaining the efficiency required for real-time moderation.

## 2 Literature Review

Research in computational mental-health natural language processing has evolved from early sentiment-based analyses to more nuanced models capable of identifying complex user intentions and mental-health signals. Initial efforts such as the CLPsych Shared Tasks (2) and the eRisk Challenge (3) established benchmark datasets for early risk detection, suicide ideation classification, and distress monitoring from social media platforms including Reddit and Twitter. These shared tasks demonstrated the limitations of traditional machine learning pipelines based on TF–IDF features and linear classifiers, particularly in capturing the contextual subtleties of mental-health discourse.

The introduction of transformer-based language models, notably BERT and RoBERTa, quickly improved performance across intent and risk-detection tasks. However, subsequent work highlighted the importance of domain adaptation: models pretrained exclusively on generic corpora often struggled to generalize to mental-health narratives. This led to the development of domain-specific encoders such as MentalBERT and DistilMentalBERT (4), which were pretrained on large-scale Reddit mental-health communities and showed consistent gains across emotion classification, self-harm detection, and psychosocial risk assessment.

Recent research has investigated augmentation and prompting-based strategies to further mitigate data scarcity, a recurring challenge in mental-health NLP. Liyanage et al. (1) proposed a prompt-based augmentation framework for classifying Wellness Dimensions, leveraging GPT-3.5 and related generative models to expand annotated datasets. Their approach produced notable improvements in macro-F1 and Matthews correlation coefficient compared to traditional augmentation methods such as back-translation and EDA-style perturbations, demonstrating the effectiveness of generative augmentation in low-resource clinical domains.

Another line of work has focused on ensuring safety, interpretability, and ethical deployment. Studies such as Valdez et al. (2023) and Harrigian et al. (2022) emphasize the risks of biased predictions, misclassification of high-risk content, and the necessity of transparent model behavior when used in clinical or support settings. Parallel to these developments, retrieval-augmented generation (RAG) architectures (5) have emerged as a promising paradigm for grounded and context-aware response generation. By integrating external knowledge retrieval with generative models, RAG methods have been shown to improve factual consistency and reduce hallucinations—properties that are particularly important in mental-health applications requiring empathetic but safe feedback.

Parallel to these developments, retrieval-augmented generation (RAG) architectures (5) have emerged as a powerful paradigm for grounded and context-aware language generation. RAG models operate by retrieving semantically relevant evidence from an external knowledge base and conditioning the generator on this retrieved context. Lewis et al. (5) demonstrated that incorporating retrieval substantially enhances factual accuracy, reduces hallucinations, and improves task performance on knowledge-intensive benchmarks. The core intuition is that external memory acts as a dynamic information source, enabling the generator to produce responses that are better grounded in real data rather than relying solely on parametric memory. These characteristics make RAG particularly relevant for safety-critical domains such as mental health support, where responses must remain empathetic, factual, and grounded in validated counseling examples.

The effectiveness of retrieval-based methods depends heavily on the ability to perform fast and accurate similarity search over large embedding spaces. Johnson et al. (7) addressed this challenge with FAISS, a GPU-optimized library designed for billion-scale vector retrieval. FAISS introduced highly efficient indexing structures such as IVF, HNSW, and product quantization that dramatically reduce retrieval latency without compromising semantic precision. This made large-scale RAG architectures computationally feasible and remains a foundational component of modern retrieval pipelines. In systems requiring rapid retrieval from thousands of counseling conversations or clinical documents, FAISS enables low-latency nearest-neighbor search essential for maintaining responsiveness and scalability.

Building on these foundations, domain-specific RAG frameworks have demonstrated the value of retrieval-grounded generation in sensitive, high-stakes verticals. For example, Han et al. (8) proposed MedRAG, a medical evidence retrieval-augmented system designed for clinical question answering. MedRAG integrates medical knowledge retrieval with controlled generation to improve factual correctness, mitigate harmful suggestions, and provide evidence-backed reasoning. Their work highlights how RAG can be adapted to domains requiring both accuracy and safety, principles directly aligned with the objectives of mental-health support systems.

# 3 Dataset Curation and Preprocessing

The validity of any machine learning model is strictly bounded by the quality of its training data. We curated a comprehensive multi-source corpus comprising mental-health discourse from Reddit, spanning the years of 2022 to 2025.

## 3.1 Data Acquisition

We aggregated data from five primary resources:

- *Reddit Mental Health Dataset (Zenodo)* – A foundational corpus containing posts from major support subreddits.
- *Sentiment Analysis for Mental Health (Kaggle)* – A dataset labeled for specific disorders providing ground truth for diagnostic categories.
- *Reddit Depression Cleaned (HuggingFace)* – A pre-cleaned dataset, free from typical social media noise.
- *Wellness Dimensions (Garg et al.)* – Used primarily for our initial augmentation experiments to understand wellness categories.
- *Mental Disorders Identification Reddit NLP (Kaggle)* – An unlabeled corpus used for domain enrichment during training phase.

We implemented a rigorous preprocessing pipeline to normalize the input data before it reached any annotation or modeling stage (Concatenation, Lowercasing, URL Removal, Emoji Removal, Anonymization).

## 3.2 Annotation Process

A subset of 500 posts was selected for Gold Standard manual annotation, using Label Studio interface.

- *Intent (Multi-label):* A user often has multiple goals. We identified nine distinct categories – Cause of Distress, Miscellaneous, Mental Distress, Positive Coping, Critical Risk, Progress Update, Seeking Help, Maladaptive Coping and Mood Tracking.
- *Concern Level:* (Low, Medium, High) indicating distress and urgency.

To scale from 500 posts to 6000 posts, we trained a preliminary classifier on the Gold Standard. We ran this model on remaining unlabeled data and accepted predictions with a confidence score of >= 0.85 as pseudo labels.

## 3.3 Data Augmentation and Stratification Strategy

To mitigate the scarcity of labeled high-risk data, we iteratively developed an augmentation pipeline to enhance model robustness. Initially, we employed a rule-based lexical substitution strategy using Python regular expressions and the NLTK WordNet interface. This method performed stochastic synonym substitution, specifically targeting adjectives and verbs to introduce linguistic variation. However, this approach proved brittle; it failed to account for the idiosyncratic flow of Reddit speech and struggled to preserve semantic coherence in long-form narratives, particularly when the core meaning was situated at the middle or end of the post.

To overcome these limitations, we transitioned to a generative augmentation pipeline utilizing **Facebook BART**. By finetuning BART on our seed dataset, we leveraged its sequence-to-sequence capabilities to generate semantically equivalent paraphrases. Unlike the regex approach, BART operates at the semantic level, effectively preserving the boundaries between intent categories. This generative process produces high-quality training samples that encourage the downstream classifier to learn latent features and underlying sentiment, rather than overfitting to specific keywords.

The final augmented corpus was partitioned into Training, Validation, and Test sets. We employed **stratified sampling** based on the *Concern* label to ensure class consistency across splits. This stratification was critical for preventing the model from biasing toward the majority class (Low Concern) and ensuring rigorous evaluation on high-risk instances.

## 4 Methodology: Model Architectures

We developed a modular pipeline using the Hugging Face Transformers library to parse mental health discourse. To maximize interpretability and minimize interference between tasks, we trained separate classifiers for **Intent Detection** and **Concern Assessment** rather than employing a single multitask architecture. Both classifiers utilize a sigmoid-activated binary cross-entropy objective function, treating each class prediction as an independent probability. This approach is particularly critical for intent classification, where a single user post may simultaneously exhibit multiple signals, such as "Seeking Help" and "Venting."

### 4.1 Hierarchical Model Evaluation

Our experimental strategy employed a three-tiered hierarchy designed to evaluate the trade-offs between interpretability, computational efficiency, and semantic depth.

We established a transparent baseline using **MiniLM coupled with Logistic Regression**. This architecture utilizes `all-MiniLM-L6-v2`, a sentence transformer distilled from BERT, to map input texts into a fixed 384-dimensional dense vector space. By freezing these embeddings and training a linear Logistic Regression classifier on top, we ensured total transparency. Unlike deep neural networks, this configuration allows us to inspect the specific weights assigned to each embedding dimension, providing a clear explanatory path for why a specific post was classified as a certain intent.

Moving beyond the baseline, we implemented **DistilRoBERTa** as our candidate for production deployment. This distilled version of RoBERTa features 6 layers and is approximately 40% smaller and 60% faster than its teacher model while retaining 97% of the performance. Finally, to test the upper limits of classification capability, we employed **RoBERTa-Large**. As a "powerhouse" model with 24 layers, 1024 hidden units, and 355 million parameters, it possesses the deep syntactic understanding necessary to disentangle the complex and often ambiguous nuances of mental health intents.

### 4.2 Parameter-Efficient Fine-Tuning (LoRA)

To facilitate the training of these large language models on resource-constrained hardware, we utilized Low-Rank Adaptation (LoRA). Standard fine-tuning requires updating all model parameters, which is computationally expensive. Instead, LoRA freezes the pre-trained model weights and injects trainable **rank decomposition matrices** into the layers. This technique enabled us to reduce GPU memory usage by approximately 60%, confirming the system's scalability.

We applied specific LoRA configurations to optimize each model. For DistilRoBERTa, we used a rank $r = 64$ and alpha $\alpha = 128$. For the larger RoBERTa-Large model, we applied a stricter rank $r = 16$, $\alpha = 32$, and a dropout rate of $0.1$.

### 4.3 Training Configuration

To ensure a rigorous and fair comparison across all three tiers, we adhered to a standardized training rig. All models were optimized using the AdamW optimizer with a learning rate of $1e{-}4$, a batch size of 8, and trained for 8 epochs. Performance was evaluated using Macro-F1 and Micro-F1 scores to account for the class imbalances inherent in mental health datasets.

## 5 Experiments and Results

### 5.1 Results

| Model | Macro F1 (Regex) | Macro F1 (BART) |
|---|---|---|
| MiniLM | 0.58 | 0.72 |
| DistilBERT | 0.64 | 0.72 |
| RoBERTa + LoRA | 0.70 | 0.77 |

Table 1: Intent classification performance using Regex-derived and BART-derived labels.

| Model | Macro F1 (Regex) | Macro F1 (BART) |
|---|---|---|
| MiniLM | 0.69 | 0.70 |
| DistilBERT | 0.78 | 0.74 |
| RoBERTa + LoRA | 0.76 | 0.81 |

Table 2: Concern classification performance using Regex-derived and BART-derived labels.

### 5.2 Observations

Our analysis began with models trained on the original Regex-constructed dataset, which provided a keyword-driven baseline for both intent and concern classification. While this dataset enabled initial benchmarking, its limited contextual coverage motivated the development of a second, more semantically grounded dataset generated through BART tagging. The BART-based method labels posts using contextual inference rather than surface-level patterns, producing richer supervision signals for downstream classifiers.

Introducing this BART-tagged dataset led to a consistent improvement across all benchmarking models, yielding an average 5% increase in F1 scores. The strongest gains appear in intent classification, where semantic categories such as Cause of Distress, Maladaptive Coping, and Mental Distress often overlap lexically; contextual tagging helped clarify these boundaries. Using this enriched dataset, RoBERTa-Large + LoRA achieved the highest intent performance (0.77 Macro-F1), showing superior discrimination across nine overlapping intent labels.

For concern-level classification, RoBERTa + LoRA attained the best results (0.81 F1) on the BART-tagged dataset, particularly excelling in distinguishing Low versus Medium distress.

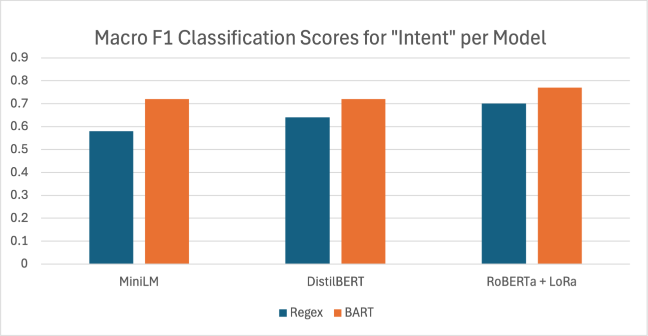


Figure 1: Intent Classification Scores

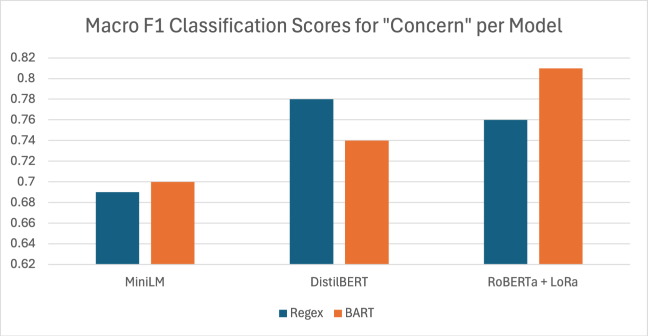


Figure 2: Concern Classification Scores

The primary difficulty arises between Medium and High categories, where emotional tone and narrative intensity often converge. By contrast, DistilRoBERTa showed a noticeable drop in concern F1 when moving from regex-tagged to BART-tagged data, suggesting that smaller models struggle to generalize across fine-grained distress levels.

Despite being lightweight, the MiniLM + Logistic Regression baseline remained competitive, achieving 0.72 Intent F1 and 0.70 Concern F1, comparable to DistilRoBERTa on intent prediction. Its stable feature separability and interpretability make it a useful baseline for transparent analysis. Finally, across all experiments, LoRA-based fine-tuning enabled rapid convergence while reducing GPU memory usage by approximately 60

To ensure fair evaluation, we applied a dynamic split by concern level, enforcing balanced representation across Low, Medium, and High distress categories. This mitigates bias toward more frequent low-severity samples and improves robustness across all model families.

# 6 RAG: Retrieval-Augmented Support Generation

To extend the intent and concern classifiers into actionable moderator assistance, we developed a Retrieval-Augmented Generation (RAG) module that generates grounded, empathetic, and safety-aligned suggestions. The pipeline is designed specifically for mental-health discourse, where hallucination control, crisis compliance, and grounding are critical.

## 6.1 Knowledge Base Construction

We curated a repository of verified peer-support responses using the MentalChat16K dataset. Each post–reply pair was cleaned, filtered for safety, and encoded using `all-MiniLM-L6-v2`. These 384-dimensional embeddings were indexed using FAISS (HNSW), enabling fast similarity search and scalable retrieval.

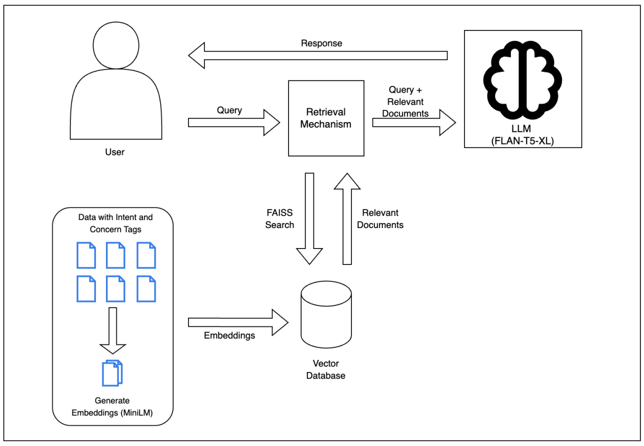


Figure 3: RAG Architecture

## 6.2 Retrieval Mechanism

For each new Reddit post, we encode the text using the same MiniLM model employed during knowledge base construction. Its embedding is used to query the FAISS index, retrieving the top-$K$ semantically relevant peer-support examples. To ensure that only high-quality and safe content is forwarded to the generator, we apply a two-stage filtering process. First, a semantic similarity threshold removes weak or loosely related matches that do not contribute meaningful grounding. Second, a safety filter eliminates any replies containing instructions, explicit methods, or other forms of unsafe language. After these filtering steps, the top five to seven remaining snippets constitute the grounding context supplied to the generation module.

## 6.3 Generation with Flan-T5-XL

We use `Flan-T5-XL` as the generator, formatting each set of retrieved snippets into a controlled prompt that clearly defines the model's output behavior. Instead of copying text verbatim, the model is instructed to paraphrase supportive patterns from the retrieved examples while avoiding any first-person roleplay or personal anecdotes. The prompt also constrains the response length to a concise 3–5 sentences and requires the inclusion of citation tags such as [S1] to ensure explicit grounding.

This structured prompting strategy reduces hallucination, maintains stylistic consistency, and guides the model toward producing evidence-based, peer-support style responses aligned with the retrieved examples.

## 6.4 Crisis Handling

We apply a custom crisis detector to every post. If the post contains self-harm or suicidal intent, the system bypasses generation and directly returns an emergency-safe crisis footer referencing the 988 Lifeline and international resources. For noncritical posts, the generated text is retained and citations are added if missing.

### 6.5 Evaluation Framework

We developed a four-part evaluation framework to systematically assess the quality of RAG-generated outputs. The first component, relevance, measures how well the reply aligns with the user's post using MiniLM-based semantic similarity. To avoid unfair penalties on emotionally expressive text—common in peer-support settings—we apply a smoothing function that reduces sensitivity to variations in affective language. The second component, grounding, evaluates how effectively the model incorporates retrieved evidence. This is captured through a hybrid score that blends lexical overlap with the snippet content and semantic similarity to the underlying retrieved examples.

Safety is treated as a standalone requirement: each reply is checked for the presence of harmful, directive, or otherwise unsafe phrasing, resulting in a binary safety indicator. Finally, we assess crisis coverage, ensuring that whenever a post expresses suicidal intent, the model's response includes an appropriate crisis-support footer. Together, these four dimensions provide a comprehensive view of both the linguistic quality and the ethical reliability of the generated responses.

| **Evaluation Component** | **Weight** |
|---|---|
| Grounding Quality | 0.45 |
| Relevance (Semantic Similarity) | 0.30 |
| Safety Compliance | 0.20 |
| Crisis Footer Compliance | 0.05 |
| **Mean Final Score** | **0.73** |

Table 3: RAG Quality Evaluation Summary.

All outputs passed the safety filter, and crisis-flagged posts consistently triggered emergency-resource messages.

Overall, the RAG module demonstrates that retrieval-grounded generation can produce concise, empathetic, and safety-compliant support suggestions while minimizing hallucination—an essential requirement for mental-health moderation workflows.

### 6.6 Limitations

- **Model:** Although Flan-T5-XL performs well in controlled settings, it sometimes produces generic or emotionally flat responses and struggles with deep contextual nuance compared to more recent instruction-tuned models (e.g., Llama-3, Mistral-Instruct). Its relatively small input window (512–1024 tokens) restricts how much retrieved evidence can be incorporated, and the lack of domain-specific mental-health fine-tuning limits sensitivity to subtle clinical cues.
- **Evaluation:** While the evaluation framework measures relevance, grounding, safety, and crisis-footer compliance, these automated metrics cannot assess therapeutic appropriateness, empathetic depth, cultural nuance, or clinical suitability. Reliable assessment of supportive communication ultimately requires expert human-in-the-loop evaluation.
- **Knowledge Base:** Expanding the knowledge base at scale requires large volumes of professionally authored or clinician-reviewed counseling data. Such resources are difficult to obtain due to privacy restrictions, licensing barriers, and ethical considerations, making long-term scalability dependent on access to trusted, expert-verified datasets.

## 7 Future Work

An important direction for future development is the integration of structured logging and monitoring for the retrieval augmented generation module, especially when the system produces suggestions for posts classified as high or immediate risk. These logs can be used to create an alerting pipeline for moderators, allowing them to review potentially urgent posts in real time. This approach preserves a human in the loop workflow where moderators verify the model output before any emergency action is taken. Such a design reduces the risk of false alarms and ensures that escalation decisions remain under human oversight, which is essential when interacting with sensitive mental health content.

A second opportunity for extension is the deployment of the system as a platform level tool for large online communities such as Reddit. The model could automatically analyze posts submitted to mental health subcommunities, assign intent and concern labels, and generate supportive responses through the retrieval augmented generation component. This integration would make the classifier and response generator accessible at scale while maintaining the transparency and interpretability goals of the project. Future iterations may incorporate adaptive learning from moderator feedback, continual alignment checks, and expanded retrieval sources in order to further strengthen the safety and reliability of the system.

## 8 References

[1] C. Liyanage, M. Garg, V. Mago, and S. Sohn, "Augmenting Reddit Posts to Determine Wellness Dimensions impacting Mental Health," in Proc. Conf. Assoc. Comput. Linguist. BioNLP Workshop, Jul. 2023, pp. 306–312. DOI: 10.18653/v1/2023.bionlp-1.27.

[2] P. Resnik et al., "The CLPsych Shared Task Series: Applying NLP to Mental Health," CLPsych Workshop, 2015–2021.

[3] D. Losada et al., "eRisk 2018: Early Risk Prediction on the Internet," CLEF Labs, 2018.

[4] S. Ji et al., "MentalBERT and DistilMentalBERT: Domain-specific Pretrained Transformers for Mental Health Texts," EMNLP, 2022.

[5] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.

[6] D. Valdez et al., "Ethical NLP for Mental Health: Balancing Privacy, Safety, and Interpretability," ACL, 2023.

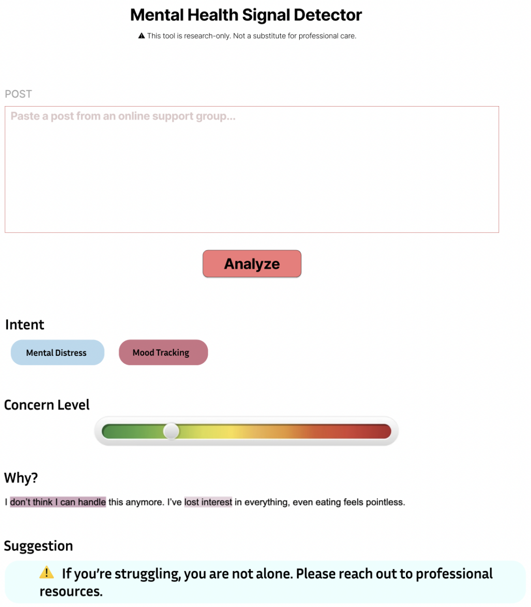
[7] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," IEEE Transactions on Big Data, 2019.

[8] S. Han et al., "MedRAG: Medical Evidence Retrieval-Augmented Generation," arXiv preprint arXiv:2304.09743, 2023.

## 9 Appendix

## A A Demonstration of the Mental Health Signal Detector

This appendix presents a demonstration of the Mental Health Signal Detector using the interactive interface developed for this project. The interface shows the complete processing flow for a post submitted by a user in an online peer support community. When a post is entered into the text field, the system produces three outputs. The first output is the predicted intent, where the model identifies one or more categories such as mental distress, mood tracking, positive coping, seeking help, or related expressions of emotional state. These intents are displayed as clearly separated tags to preserve interpretability. The second output is the predicted concern level, which is shown through a continuous color based slider that ranges from low to high and provides an immediate visual indication of the urgency associated with the post. The third output is an automatically generated suggestion delivered through a retrieval augmented generation module. This suggestion is designed to provide short, empathetic, and safety aligned guidance, especially when the concern level is elevated. The interface also highlights short text segments that influence model predictions to support transparent and responsible use without overwhelming the reader with technical detail. Together, these elements illustrate how intent classification, concern estimation, and retrieval grounded guidance can be integrated into a single user flow that supports moderators and users in understanding and responding to emotionally sensitive content.



Mental Health Signal Detector

▲ This tool is research-only. Not a substitute for professional care.

Intent

Concern Level

Why?

I don't think I can handle this anymore. I've lost interest in everything, even eating feels pointless.

Suggestion

⚠ If you're struggling, you are not alone. Please reach out to professional resources.

Figure 4: Prototype interface of the Mental Health Signal Detector.