Explainable Depression Detection in Clinical Interviews with Personalized Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Depression is a widespread mental health disorder, and clinical interviews are the gold standard for assessment. However, their reliance on scarce professionals highlights the need for automated detection. Current systems mainly employ black-box neural networks, which lack interpretability, which is crucial in mental health contexts. Some attempts to improve interpretability use post-hoc LLM generation but suffer from hallucination. To address these limitations, we propose RED, a Retrieval-augmented generation framework for Explainable depression Detection. RED retrieves evidence from clinical interview transcripts, providing explanations for predictions. Traditional query-based retrieval systems use a one-size-fits-all approach, which may not be optimal for depression detection, as user backgrounds and situations vary. We introduce a personalized query generation module that combines standard queries with user-specific background inferred by LLMs, tailoring retrieval to individual contexts. Additionally, to enhance LLM performance in social intelligence, we augment LLMs by retrieving relevant knowledge from a social intelligence datastore using an event-centric retriever. Experimental results on the real-world benchmark demonstrate RED's effectiveness compared to neural networks and LLM-based baselines.

1 Introduction

Depression is one of the most prevalent mental health disorders, affecting millions of individuals worldwide (Fava and Kendler, 2000). Timely detection is crucial for effective intervention, yet traditional assessment methods, such as clinical interviews, rely heavily on trained professionals, which are in short supply (Kroenke et al., 2001). As a result, there is an increasing need for automated systems capable of accurately detecting depression from patient interactions, enabling faster and more widespread diagnosis (Islam et al., 2018; Orabi et al., 2018; Chen et al., 2024).

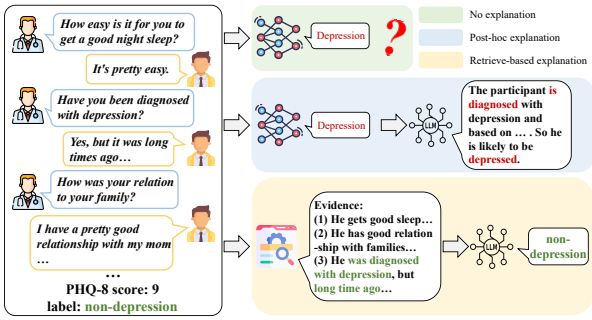

Figure 1: Comparison between different depression detection methods. Most of the methods focus on improving performance while ignoring the explanation. Some work tries to generate post-hoc explanations with LLMs while suffering from the hallucination. Our work employs a RAG-based framework to retrieve the supporting evidence from dialogue, which serves as the explanations for the predictions.

Most automated depression detection methods currently focus on enhancing system performance using various approaches. Previous work has concentrated on aggregating word representations for prediction (Mallol-Ragolta et al., 2019; Burdisso et al., 2023), or further incorporating affective and mental health lexicons (Xezonaki et al., 2020; Villatoro-Tello et al., 2021). Some studies have explored modalities beyond text, such as audio (Ma et al., 2016; Sardari et al., 2022), or have integrated multimodal data (Al Hanai et al., 2018; Wu et al., 2022). However, in health-related tasks, precision is not the sole priority; it is also crucial to understand the rationale behind the system's predictions to make the system more transparent and reliable. To achieve this, some studies have utilized the internal states of neural models, such as attention scores (Zogan et al., 2022), though these explanations often fall short of human interpretability. In response, some researchers have leveraged the generative capabilities of large language models

(LLMs) to create post-hoc explanations based on the system's predictions (Wang et al., 2024).

As shown in Figure 1, most previous work overlooks the interpretability of the system, making its predictions difficult to understand and less reliable (Mallol-Ragolta et al., 2019; Burdisso et al., 2023). Some studies have attempted to generate post-hoc explanations using large language models (LLMs), but these approaches are often plagued by the issue of hallucinated generation (Wang et al., 2024). To address these challenges, we propose employing the Retrieval-Augmented Generation (RAG) framework for explainable depression detection. The RAG framework combines a retriever model with LLMs to improve the LLM's ability to handle content beyond its input window and to update its knowledge (Lewis et al., 2020; Gao et al., 2023). With the RAG framework, crucial information from the interview dialogue is retrieved and serves as supporting evidence for the LLM's prediction. Furthermore, the retrieval process helps filter out noisy or irrelevant information that could negatively impact the prediction. Since the evidence is directly retrieved from the interview text, it is both human-understandable and free from hallucinations.

In this paper, we propose RED, a Retrieval-augmented generation framework for Explainable depression Detection. RED retrieves relevant evidence from clinical interview transcripts and uses this information as an explanation for its predictions. This retrieval-based approach ensures that the explanations are grounded in the actual content of the interview, enhancing transparency. Since depression detection is a personalized task where participants' backgrounds can vary, the traditional approach of using a single, unified query for all users may lead to suboptimal results. To address this, we introduce a personal query generation module in RED, which tailors the basic query to each individual based on their profile, inferred from the LLM. This customization enables more accurate and context-sensitive predictions. While LLMs have proven effective across many tasks due to their extensive world knowledge, they often lack domain-specific knowledge and fall short of social intelligence (Wang et al., 2023; Hou et al., 2024; Liu et al., 2024). To mitigate this, we enhance the social intelligence of LLMs by retrieving external knowledge from a social intelligence knowledge base using an event-centric retriever. Experimental results on a real-world depression detection benchmark demonstrate that RED outperforms both neural network-based and LLM-based methods, highlighting its effectiveness.

The contributions of this paper can be summarized as:

- **New framework**: We propose to perform explainable derepression detection based RAG framework with personal retrieve process;
- **LLM social intelligence enhancement**: A novel module that enhances the social intelligence of LLM based on event-centric retrieval is proposed;
- **Empirical Performance**: Experimental results demonstrate the effectiveness of our approach compared to both neural network-based and LLM-based baselines.

2 Related Work

2.1 Depreesion Detection

Depression detection is challenging due to its subtle nature, with traditional methods relying on clinical interviews or social media data (Gratch et al., 2014; Burdisso et al., 2020; Salas-Zárate et al., 2022). Recent approaches focus on multi-modal data from interviews, combining text, audio, and video for better accuracy (Gratch et al., 2014; Burdisso et al., 2020; Salas-Zárate et al., 2022). These methods aggregate features at various levels (word or utterance) to capture more nuanced signs of depression. Early risk detection is also gaining traction, using techniques like incremental classifiers and risk window-based methods to predict depression before full symptoms emerge (Burdisso et al., 2019a,b; Sadeque et al., 2018). These approaches enable timely interventions by detecting consistent patterns over time. Recent advancements also incorporate LLMs, which are fine-tuned on mental health datasets to capture complex linguistic and psychological cues. By combining LLMs with multimodal data, these methods show promise in improving both depression detection and early intervention (An et al., 2020; Yoon et al., 2022).

2.2 Retrievel Augementated Generation

Retrieval-Augmented Generation (RAG) enhances language models (LMs) by incorporating retrieved text passages into the input, leading to significant improvements in knowledge-intensive tasks (Guu et al., 2020; Lewis et al., 2020). Recent advancements in RAG techniques have focused on

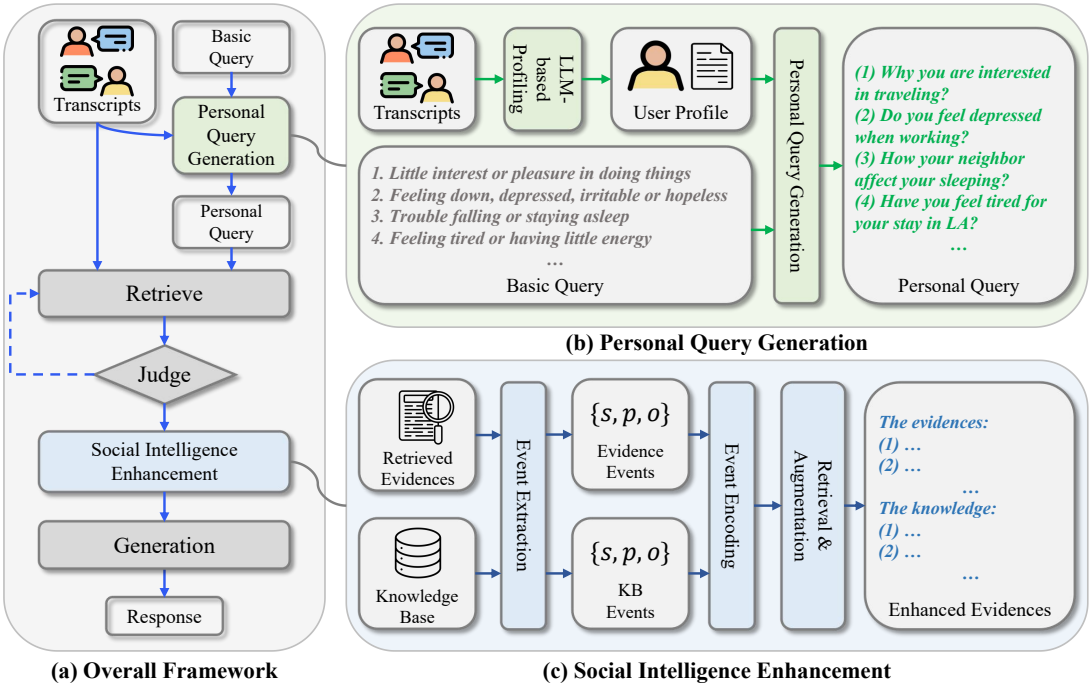

Figure 2: Overview of RED, which consists of (a) The adaptive RAG framework with two important modules, (b) the Personal Query Generation module, and (c) the Social Intelligence Enhancement module.

instruction-tuning LMs with a fixed number of retrieved passages or jointly pre-training a retriever and LM followed by few-shot fine-tuning (Luo et al., 2023; Izacard et al., 2022). Some approaches adaptively retrieve passages during generation (Jiang et al., 2023), while others, like Schick et al. (2023), train LMs to generate API calls for named entities. However, these improvements often come with trade-offs in runtime efficiency, robustness, and contextual relevance (Mallen et al., 2023; Shi et al., 2023). To address these challenges, recent work introduces methods like SELF-RAG, which enables on-demand retrieval and filters out irrelevant passages through self-reflection, enhancing robustness and control (Lin et al., 2024; Yoran et al., 2024). SELF-RAG (Asai et al., 2023) also evaluates the factuality and quality of the generated output without relying on external models during inference, making it more efficient and customizable. Additionally, other concurrent RAG methods, such as LATS (Zhou et al., 2023), explore ways to improve retrieval for specific tasks like question answering through tree search.

3 Method

3.1 Overall Framework

As shown in Figure 2, RED utilizes an adaptive RAG framework for depression detection, using retrieved chunks as explanations. First, the personal query generation module customizes the basic query based on the inferred user profile. Then, the system retrieves relevant evidence for depression prediction based on the personalized query and transcription, while the judgment module produces a stop signal. The retrieved evidence is further enhanced with knowledge from the social intelligence knowledge base through an additional retrieval process. Finally, the LLM generates the response using the enhanced evidence.

We will discuss the RAG framework, the personal query generation module, and the social intelligence enhancement module in detail.

3.2 Explainable Depression Detection with Adaptive RAG

Previous approaches use LLMs to generate post-hoc explanations for depression detection, which often suffer from hallucination issues. In this paper, we use Retrieval Augmented Generation (RAG) to first retrieve relevant dialogue snippets from the dialogue set as evidence, then integrate this evidence into the LLM prompts for predictions. This approach ensures the system uses reliable, relevant snippets, avoiding irrelevant content and improving precision. Additionally, the retrieved snippets serve as explanations for the system's output, enhancing interpretability. As evidence requirements may

Algorithm 1: RED inference process

Data: Dialogue D, PHQ-8 aspect set $A = \{a\}$,
Basic query set $Q = \{q_a\}$, Knowledge base K
Result: the predicted label $\hat{y}$
$re =$ TRUE
$Q^{(u)} =$ PQ-Gen(Q, D) ▷ Section 3.3
for $a \in A$ **do**
  $D_a = D$
  while $re =$ *TRUE* **do**
    $e =$ Retrieve$(q_a^{(u)}, D_a)$ ▷ Retrieve
    $E_a = E_a \cup e$
    $D_a = D_a / e$
    $re =$ Judge$(q_a^{(u)}, E_a)$ ▷ Judge
$E = \bigcup_{a \in A} E_a$
$E_s =$ SI-Enh(E, K) ▷ Section 3.4
$\hat{y} =$ Geneartion(E_s) ▷ Geneartion

vary across users, we employ an adaptive RAG framework that allows the system to determine when to stop retrieving automatically.

The detailed inference process of RED is shown in Algorithm 1. The inputs include the dialogue D, the PHQ-8 aspect set $A = a$, the basic query set $Q = \{q_a\}$, and the knowledge base K. The output is the predicted depression label $\hat{y} = 0, 1$. First, RED tailors the basic query $Q = \{q_a\}$ into a personal query $Q^{(u)} =$ PQ-Gen(Q, D) using the personal query generation module, PQ-Gen(Q, D), detailed in Section 3.3. Next, the personal query $Q^{(u)}$ is used to retrieve relevant evidence e from D, based on the aspects $A = a$ aligned with the PHQ-8 questionnaire. Then, the social intelligence enhancement module, SI-Enh(E, K), augments the evidence E with knowledge from the knowledge base K, as explained in Section 3.4. Finally, the LLM generates the final response $\hat{y}$ based on the enhanced evidence E_s.

Retrieve The standard process for depression detection is based on the PHQ-8 (Patient Health Questionnaire-8) (Kroenke et al., 2001), where the participant self-evaluates on 8 questions addressing various aspects of depression, such as interest in activities and issues with movement or speech. Each question is scored from 0 (Not at all) to 3 (Nearly every day), resulting in a total score from 0 to 24. A score above 10 typically indicates depression.

Building on this process, we propose retrieving evidence based on different aspects. For each aspect a, a personal query $q_a^{(u)}$ is used to retrieve evidence e from the dialogue set D_a, forming the evidence set E_a. RAG typically uses sparse retrievers (e.g., BM25) and dense retrievers. In this work, we implement a dense retriever based on GPT embedding model [1] with L2 similarity.

$$\begin{aligned} \boldsymbol{q} &= \text{BERT}(q) \\ \boldsymbol{d}_i &= \text{BERT}(d_i) \\ e &= \text{Top-1}(\{sim(\boldsymbol{q}, \boldsymbol{d}_i)\}) \end{aligned} \tag{1}$$

where q denotes the query, $d_i \in D_a$ denotes the i-th dialogue snippet, $sim(\cdot, \cdot)$ denotes the cosine similarity. At each iteration, the top-1 result is returned, and both the evidence and dialogue sets are updated by $E_a = E_a \cup e$ and $D_a = D_a / e$.

Judge Since dialogue can shift to different topics, a one-time retrieval may be insufficient for judgment. Moreover, determining a specific threshold for retrieval is challenging. To address this, we propose a judgment module that allows the system to determine when to stop retrieving adaptively. The judgment model is a binary classification model, which can be implemented as either a supervised neural network or an LLM agent. In this paper, we implement it as an LLM agent that takes the retrieved evidence set E_a and the personal query $q_a^{(u)}$ as inputs:

$$re = \text{LLM}(q_a^{(u)}, E_a | p_j) \tag{2}$$

where re is the retrieval indicator and p_j is the judgment prompt. The full prompt can be found in Appendix A.1.

Generation The final response is generated using the enhanced evidence set E_s:

$$\hat{y} = \text{LLM}(E_s | p_g) \tag{3}$$

where p_g is the depression detection task prompt. The full prompt is provided in Appendix A.1.

3.3 Personal Query Generation

The depression diagnosis interview process is highly personalized and can vary across users (Goldman et al., 1999). Thus, using a single query for all participants may lead to suboptimal results. To address this, we propose tailoring the basic query to each participant's background, creating a personal query. Inspired by profile-augmented generation in personalized LLMs (Richardson et al., 2023), we first infer the user profile from dialogues with an LLM agent for participant u:

$$d_u = \text{LLM}(D | p_d) \tag{4}$$

[1] https://platform.openai.com/docs/guides/embeddings

where p_d is the user profiling prompt, with the full prompt available in Appendix A.1. Next, the personal query $q_a^{(u)}$ for aspect a is generated from the basic query q_a:

$$q_a^{(u)} = \text{LLM}(q_a, d_u | q_p) \tag{5}$$

where q_p denotes the personal query generation prompt, and the full prompt can be found in Appendix A.1 along with the basic queries template.

3.4 Soical Inerllegence Enhancement

Although LLMs perform well on various tasks with extensive knowledge, they still lack social intelligence and psychological understanding. To enhance LLM judgment, we propose retrieving relevant knowledge from a psychological knowledge base to augment the evidence (Wu et al., 2024).

However, retrieving relevant knowledge from dialogue can be challenging due to its rich and noisy content. Inspired by event-centric sentiment analysis (Zhou et al., 2021), we treat key elements in the dialogue as events that happened to the participant. This leads to an event-centric retrieval approach. By extracting key events from the dialogue snippets, we can focus on relevant information for more accurate retrieval.

To extract events from text, one could use a supervised event extraction model or an LLM agent. In this paper, we employ the LLM agent to extract event triplets s, p, o from text t, where s is the subject, p is the predicate and o is the object:

$$\{s, p, o\} = \text{LLM}(t | p_e) \tag{6}$$

where p_e is the event extraction prompt. We perform event extraction for both the dialogue sentences and the keys in the knowledge base to ensure alignment.

With the extracted events, we perform event representation learning using the event encoder from MORE-CL (Zhang et al., 2023), where event triplets are projected into a Gaussian embedding space, and similarity is calculated using symmetric KL-divergence. Formally, the knowledge base is represented as $K = (k_i, v_i)$, where k represents the key and v represents the value.

$$\begin{aligned} (\boldsymbol{\mu}, \boldsymbol{\sigma}) &= \text{E-encoder}(\{s, p, o\}) \\ \{s\} &= \text{Top-k}(\text{KL}(g_i, g_j) + \text{KL}(g_j, g_i)) \end{aligned} \tag{7}$$

where $\{s\}$ are retrieved knowledge pairs.

4 Experimental Settings

4.1 Datasets

We conduct experiments on an available corpus for clinical depression detection: the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WoZ) (Gratch et al., 2014), which is a widely utilized English-language dataset comprising interviews from 189 participants, with data available in the form of transcripts, audio recordings, and videos. After the interaction, participants are asked to complete the PHQ-8 questionnaire (Kroenke et al., 2009), which assesses eight specific symptoms related to depression. These symptoms include loss of interest, feelings of depression, sleep disturbances, fatigue, loss of appetite, feelings of failure, lack of concentration, and reduced movement. Participants scoring 10 or higher are classified as depressed, while those with scores below 10 are classified as control. Detailed statistic of the dataset can be found in Appendix A.2. Following the prior research (Chen et al., 2024) and the specificity of our methodology, both the development and training sets are utilized for evaluation, as the labels for the test set are unavailable.

We do not employ another interview-based depression detection dataset EATD (Shen et al., 2022a) because it is not fully the clinical setting, where each participant was only asked three questions, making the dialogue content too short for retrieval.

For the social intelligence enhancement module, we employ COKE, a cognitive knowledge graph for machine theory of mind (Wu et al., 2024). COKE contains a series of cognitive chains to describe human mental activities and behavioral/affective responses in social situations. In RED, we concat *situation* and *clue* in COKE as the query for retrieval, and the rest of the elements as values. Detailed for the dataset can be found in Appendix A.2.

4.2 Implementation Details

For the implementation of RED, we employ the GPT model to generate personalized queries based on the basic query and the User Profile, which is summarized from the transcripts using `gpt-4o-2024-08-06`. In the retrieval phase, we use the `text-embedding-3-large` embeddings to encode both queries and transcripts and apply the L2 distance metric to retrieve the top K evidence (with $K = 10$ as the default). The COKE dataset, which contains rich and diverse scenes, is used

as the Knowledge Base. To extract event triplets s, p, o from text, we utilize a LLM agent. Each event triplet is then encoded into a Gaussian embedding space ($dims = 500$) using MORE-CL. The similarity between event triplets is calculated using the L2 distance metric, and the top M triplets are retrieved (with $M = 2$ as the default).

We followed previous studies (Burdisso et al., 2023) and, in addition to considering the depressed, control, and macro F1 scores, we also included precision and recall for both the depressed and control groups, resulting in a total of seven evaluation metrics. We select the checkpoint for evaluation based on macro F1 scores. The final results for comparison are the average scores of 3 runs. We run all experiments on a single NVIDIA GeForce RTX 3090 in Windows 11. For the LLM, we use `gpt-4o`,`gpt-4o-mini`,`gpt-4`, which are provided by OpenAI. The hyperparameter ranges can be found in Appendix A.3.

4.3 Baselines

To validate the effectiveness of the proposed RED for the Depression Detection task, we implement and compare NN-based methods and LLM-based methods.

NN-based method **ω-GCN** (Burdisso et al., 2023) is an approach for weighting self-connecting edges in a Graph Convolutional Network (GCN) **EATD-Fusion** (Shen et al., 2022b) is a bi-modal model that utilizes both speech characteristics and linguistic content from participants' interviews. **MFM-Att**(Fang et al., 2023) is a multimodal fusion model with a multi-level attention mechanism (MFM-Att) for depression detection, aiming to effectively extract depression-related features. **HCAG** (Niu et al., 2021) is a hierarchical Context-Aware Graph Attention Model model that utilizes the Graph Attention Network (GAT) to capture relational contextual information from both text and audio modalities. **SEGA** (Chen et al., 2024)transforms clinical interviews into a directed acyclic graph and enhances it with principle-guided data augmentation using large language models (LLMs)

LLM-based method **Direct Prompt** is a prompt-learning method designed to guide large language models (LLMs) in making judgments about depression. **Naive RAG** is a technique that integrates the Retrieval-Augmented Generation (RAG) framework with LLMs. It uses a retriever to search for relevant evidence from a knowledge base or dataset, which is then fed into an LLM to make judgments or generate appropriate responses. **Personal RAG** builds upon the previous method by enhancing the query generation process, which is now based on the user profile, ensuring more personalized and contextually relevant evidence retrieval.

The details for the implementation can be found in Appendix A.3.

	Method	Depressed	Control	Marco
NN -based	ω-GCN	78.26	89.36	83.81
	EATD-Fusion	69.57	85.11	77.34
	MFM-Att	78.57	85.71	82.14
	HCAG	76.92	86.36	81.64
	SEGA	81.48	88.37	84.93
	SEGA++	84.62	90.91	87.76
LLM -based	Direct Prompt	74.07	83.72	78.90
	Naive RAG	78.97	88.05	84.39
	Personal RAG	79.87	88.92	84.39
	RED	**87.83**	**92.17**	**90.00**

Table 1: Performance of RED and other baselines on the development set of DAIC-WoZ benchmark. The best scores are in **bold**. All LLM-based results are an average of three rounds of experiments based on GPT-4o.

5 Experimental Analysis

In this section, we present comprehensive experiments conducted on LaMP. Through an in-depth analysis of the results, we aim to address the following Research Questions (**RQs**):

- **RQ1**: How does RED perform compared to baseline models in a standard setting?
- **RQ2**: What impact do different architectural structures and components have on model performance?
- **RQ3**: How effective RED is in terms of explanation extraction?
- **RQ4**: How does RED perform in qualitative evaluations?

5.1 Main Results

To answer **RQ1**, we compare the performance of RED with baseline models on the DAIC-WoZ development set. The results, shown in Table 1, demonstrate that RED outperforms all baselines. We observe the following:

Method	Depressed			Control			Marco
	Precision	Recall	F-1	Precision	Recall	F-1	F-1
Direct Prompt (GPT-4)	55.24	**79.36**	65.14	89.40	73.00	80.37	72.75
Direct Prompt (GPT-4o-mini)	57.77	73.81	**73.81**	87.55	77.33	82.12	73.47
Direct Prompt (GPT-4o)	59.64	78.57	67.81	**89.62**	77.67	83.21	75.51
Naive RAG (GPT-4)	65.04	73.81	69.15	88.34	83.33	85.76	77.46
Naive RAG (GPT-4o-mini)	61.64	76.19	68.11	88.89	80.00	84.20	76.16
Naive RAG (GPT-4o)	68.15	73.02	70.49	88.32	85.67	86.97	78.73
Peronal RAG (GPT-4)	**69.93**	68.26	69.07	86.8	**87.00**	87.23	78.15
Peronal RAG (GPT-4o-mini)	60.66	69.05	64.45	86.21	81.00	83.48	73.96
Peronal RAG (GPT-4o)	68.98	72.22	70.56	88.09	86.33	**87.20**	**78.88**

Table 2: Performance of RED's variants on the full set of DAIC-WoZ benchmark. The best scores are in **bold**. All LLM-based results are an average of three rounds of experiments.

RED outperforms all baselines, especially for the depressed class. NN-based methods generally outperform LLM-based baselines with direct prompting, but RED improves LLM performance by incorporating personal retrieval and social intelligence, raising the macro F1 score from 78.90% to 90.00%. While F1 scores for the control class are high across all methods, indicating a tendency to classify most participants as control, RED achieves a significant gain in the depressed class by retrieving relevant evidence and enhancing the LLM with the necessary knowledge for accurate predictions.

The retrieval process is still necessary, even if the contents do not exceed the input window size. As shown in Table 1, the LLM baseline underperforms compared to NN-based methods. However, using a naive retrieval process, the LLM-based baseline improves, with macro F1 rising from 78.90% to 84.39%, while personal RAG achieves further gains. This improvement is due to the retrieval process filtering out irrelevant information, allowing the model to focus on what matters. Thus, the retrieval process remains essential, even when content doesn't exceed the input window size.

Social intelligence enhancement with calibration brings significant improvement. Personal retrieval improves performance over direct prompting but ties with the non-data-augmented SEGA (Chen et al., 2024). With the social intelligence enhancement, RED generates fine-grained depression scores for each PHQ-8 category, allowing for calibration. Since DAIC-WoZ transcripts do not cover all PHQ-8 aspects, such as appetite and movement difficulties, RED's predicted scores are generally lower than the actual scores. By combining social intelligence enhancement with calibration, and adjusting the threshold from 10 to 8, RED achieves substantial improvement, particularly in predicting the depressed class.

5.2 Ablation Studies

To answer **RQ2**, we compare RED with its variants on the combined DAIC-WoZ set, merging the training and development sets. We focus on two subquestions: (1) How does the capability of backbone LLMs affect RED's performance? (2) How does the retrieval module setting impact performance? The results in Table 2 reveal the following:

Training set is generally more difficult than the development set. As shown in Table 2, metrics for all variants on the combined set are lower than on the development set, suggesting the training set is more challenging. This is due to two factors: (1) The training data has a less clear decision boundary, with most participants' scores around 10, near the depression threshold, compared to the more extreme scores in the development set. (2) The training set contains more missing aspects; for example, more depressed participants reported issues with appetite and movement difficulties, leading to significant data gaps.

RED benefits from the improved capability of backbone LLMs. As seen in Table 2, performance improves consistently with the enhanced backbone LLMs (GPT-4 to GPT-4o-mini and GPT-4o) across all retrieval settings, highlighting the importance of LLMs' reasoning and instruction-following capabilities in depression detection.

The retrieval process brings universal improvement. Regardless of the backbone LLM, the retrieval process improves results at each stage. Variants based on naive RAG outperform those with direct prompting, and personal RAG variants outperform naive RAG, offering more precise retrieval

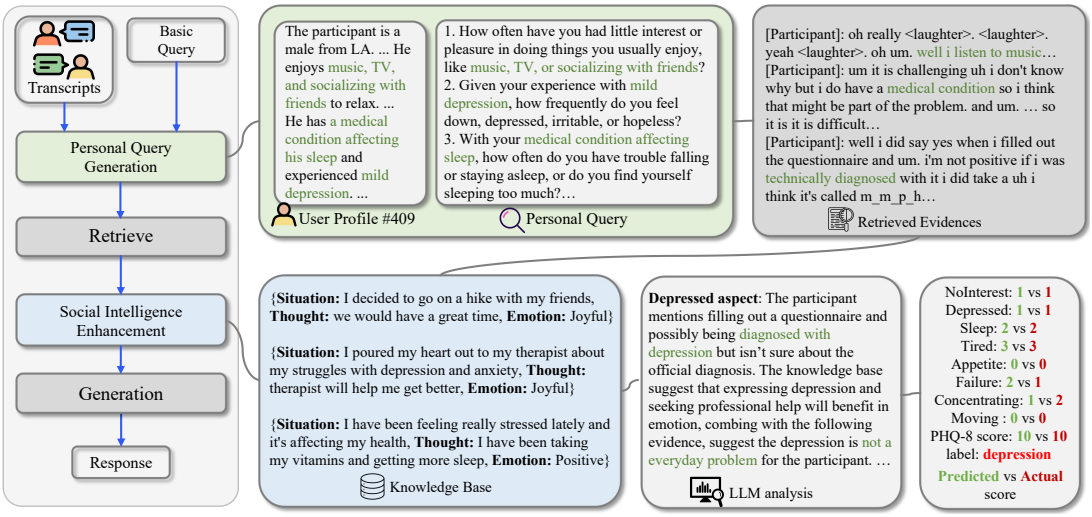

Figure 3: Case study for user #409. Texts containing personal identification information are removed. Texts in green indicate the important information for prediction, and texts in red indicate the actual scores.

	Method	Precision	Recall	F1
Non Retrieval	Direcet Prompt	30.42	59.45	40.21
	In-context Learning	47.12	62.42	56.31
Naive RAG	k=4	93.53	34.66	50.58
	k=6	91.78	49.38	64.21
	k=8	89.70	59.64	71.64
	k=10	88.30	69.02	77.48
Personal RAG	k=4	**90.62**	49.54	64.06
	k=6	88.78	63.01	73.71
	k=8	87.39	72.76	79.41
	k=10	86.78	**80.19**	**83.35**

Table 3: Performance of RED's variants on the evidence extraction on DAIC-WoZ benchmark. The best scores are in **bold**.

tailored to user backgrounds.

5.3 Explanation Extraction Analysis

To answer **RQ3**, we compare the performance of RED with other baseline models on evidence extraction. The evidence, annotated by (Agarwal et al., 2024), consists of text chunks identified by human annotators as important for depression prediction. The results, shown in Table 3, demonstrate RED's effectiveness in evidence extraction. Notably, without the retrieval system, the precision of evidence retrieval is much lower compared to the retrieval-based system, indicating that LLMs tend to generate hallucinations. With the retrieval module, both precision and recall increase significantly, and these improvements are further enhanced with the personal retrieval module. This highlights the effectiveness of personal query generation, which tailors the retrieval process to the user's background.

5.4 Case Study

To answer **RQ4**, we analyze a sample from the development set to demonstrate how RED works. Personal identification information is removed. As shown in Figure 3, this participant has a PHQ-8 score of 10, placing him at the threshold of depression, making this a challenging case to predict. However, RED successfully predicted this case, with four out of the eight aspect scores matching the ground truth. The score for the *depressed* aspect is particularly difficult, as the participant had previously been diagnosed with depression. Many systems would have incorrectly predicted this aspect with a score of 3. However, with the knowledge base enhancement and the retrieved content, RED recognized that the participant had sought help, which greatly improved his condition, leading to a predicted score of 1.

6 Conclusion

In this paper, we introduced RED, a Retrieval-Augmented Generation framework designed for explainable depression detection. By retrieving evidence from clinical interview transcripts, RED not only provides transparent explanations for its predictions but also adapts to individual user contexts through personalized query generation. Furthermore, we enhanced the RAG by retrieving social intelligence knowledge with an event-centric retriever. Experimental results on the real-world dataset validate the effectiveness of RED, demonstrating its effectiveness in explainable depression detection.

Limitations

We identify three key limitations in RED.

1) Corpus Size. Due to data collection challenges and privacy concerns, datasets for depression detection in clinical interviews are often limited in size. In our experiments, we report the average performance across multiple runs and conduct significance testing to ensure that any observed improvements are statistically valid. However, the small size of the datasets may still impact the generalizability of the results.

2) Single Modality. The DAIC dataset includes multiple modalities, such as audio and video, which could potentially enhance depression detection. Our proposed method focuses primarily on text, which is considered the most informative and widely used modality for this task. Additionally, text serves as the safest modality for protecting user privacy. That said, future work should explore the potential for explainable depression detection using multimodal data, integrating other modalities to improve the system's performance and robustness.

3) Task Format. This work focuses on explainable depression detection within clinical interviews, where dialogue snippets provide the evidence for judgment through a Retrieval-augmented Generation (RAG) framework. This setting requires interviews to be of sufficient length to provide adequate evidence for retrieval. As a result, we were unable to experiment with the EATD dataset, where each participant responded to only three questions. Additionally, the proposed method may not be easily transferred to other important settings, such as detecting early signs of depression from social media posts, due to significant differences in data structure, task format, and judgment criteria.

Ethical Impact

In developing RED for explainable depression detection, we acknowledge several potential ethical concerns related to data privacy, fairness, the role of AI in mental health, and responsible use of datasets.

1) **Data Privacy and Consent** RED utilizes clinical interview transcripts to detect depression. These datasets are crucial for building and testing the model, and we emphasize the importance of obtaining informed consent from all participants. Any personal information, such as names or identifying details, must be anonymized or removed before use to ensure privacy. Given that mental health data is particularly sensitive, stringent privacy safeguards, such as data encryption and secure handling, must be in place to protect participants from any unintended disclosures.

2) **Bias and Fairness** While RED tailors its predictions using personalized query generation based on user background, it is essential to ensure that the model does not inadvertently introduce or amplify biases. The data used for training and testing must be representative of diverse populations to avoid reinforcing stereotypes or underrepresenting specific groups, particularly those who may be vulnerable to mental health issues. We must carefully monitor the model's outputs to ensure fairness and continuous efforts should be made to detect and mitigate any bias in the system, particularly regarding sensitive demographic factors such as age, gender, or ethnicity.

3) **Role of AI in Mental Health Diagnosis** The use of RED in mental health settings should always complement, not replace, clinical expertise. While the system aims to provide valuable insights and explanations through explainable predictions, the final diagnosis and treatment decisions should remain the responsibility of qualified healthcare professionals. Using RED as an automated system for diagnosis without human oversight could lead to the misinterpretation of results, potentially harming users. The model's outputs should be viewed as recommendations or support tools, with the understanding that human judgment is essential for accurate mental health care.

4) **Responsible Dataset Use and Access** The datasets used for training RED must be handled responsibly and in compliance with all relevant ethical standards. All data must be obtained with the appropriate permissions and used strictly for research purposes. We must adhere to institutional and legal requirements when accessing and utilizing these datasets, ensuring they are not shared or disseminated without proper authorization. Further, when working with clinical datasets, it is critical to respect participant confidentiality and uphold ethical standards in all stages of data usage.

By addressing these ethical concerns, we can ensure that RED is developed and deployed in a responsible, transparent, and equitable manner, prioritizing user well-being and promoting trust in AI-driven mental health tools.

References

Navneet Agarwal, Kirill Milintsevich, Lucie Metivier, Maud Rotharmel, Gaël Dias, and Sonia Dollfus. 2024. Analyzing symptom-based depression level estimation through the prism of psychiatric expertise. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 974–983, Torino, Italia. ELRA and ICCL.

Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720.

Minghui An, Jingjing Wang, Shoushan Li, and Guodong Zhou. 2020. Multimodal topic-enriched auxiliary learning for depression detection. In *proceedings of the 28th international conference on computational linguistics*, pages 1078–1089.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Sergio Burdisso, Esaú Villatoro-Tello, Srikanth Madikeri, and Petr Motlicek. 2023. Node-weighted graph convolutional network for depression detection in transcribed clinical interviews. *arXiv preprint arXiv:2307.00920*.

Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2019a. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197.

Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2020. τ-ss3: A text classifier with dynamic n-grams for early risk detection over text streams. *Pattern Recognition Letters*, 138:130–137.

Sergio Gastón Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2019b. Unsl at erisk 2019: a unified approach for anorexia, self-harm and depression detection in social media. In *CLEF (Working Notes)*.

Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tieyun Qian, and Minlie Huang. 2024. Depression detection in clinical interviews with LLM-empowered structural element graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8181–8194, Mexico City, Mexico. Association for Computational Linguistics.

Ming Fang, Siyu Peng, Yujia Liang, Chih-Cheng Hung, and Shuhua Liu. 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561.

Maurizio Fava and Kenneth S Kendler. 2000. Major depressive disorder. *Neuron*, 28(2):335–341.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Larry S Goldman, Nancy H Nielsen, Hunter C Champion, and American Medical Association Council on Scientific Affairs. 1999. Awareness, diagnosis, and treatment of depression. *Journal of general internal medicine*, 14(9):569–580.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, volume 14, pages 3123–3128. Reykjavik.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. 2024. Entering real social world! benchmarking the theory of mind and socialization capabilities of llms from a first-person perspective. *arXiv preprint arXiv:2410.06195*.

Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. 2018. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6:1–12.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 1(2):4.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation

for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2024. RA-DIT: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. 2024. InterIntent: Investigating social intelligence of LLMs via intention understanding in an interactive game context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6718–6746, Miami, Florida, USA. Association for Computational Linguistics.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*.

Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 35–42.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn Schuller. 2019. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews.

Meng Niu, Kai Chen, Qingcai Chen, and Lufeng Yang. 2021. Hcag: A hierarchical context-aware graph attention model for depression detection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4235–4239.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 88–97.

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*.

Farig Sadeque, Dongfang Xu, and Steven Bethard. 2018. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 495–503.

Rafael Salas-Zárate, Giner Alor-Hernández, María del Pilar Salas-Zárate, Mario Andrés Paredes-Valverde, Maritza Bustos-López, and José Luis Sánchez-Cervantes. 2022. Detecting depression signs on social media: a systematic literature review. In *Healthcare*, volume 10, page 291. MDPI.

Sara Sardari, Bahareh Nakisa, Mohammed Naim Rastgoo, and Peter Eklund. 2022. Audio based depression detection using convolutional autoencoder. *Expert Systems with Applications*, 189:116076.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Ying Shen, Huiyu Yang, and Lin Lin. 2022a. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251. IEEE.

Ying Shen, Huiyu Yang, and Lin Lin. 2022b. Automatic depression detection: an emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Esaú Villatoro-Tello, Gabriela Ramírez-de-la Rosa, Daniel Gática-Pérez, Mathew Magimai.-Doss, and Héctor Jiménez-Salazar. 2021. Approximating the mental lexicon from clinical interviews as a support tool for depression detection. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 557–566.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126, St. Julians, Malta. Association for Computational Linguistics.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, Helen Meng, and Minlie Huang. 2024. COKE: A cognitive knowledge graph for machine theory of mind. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15984–16007, Bangkok, Thailand. Association for Computational Linguistics.

Wen Wu, Mengyue Wu, and Kai Yu. 2022. Climate and weather: Inspecting depression detection via emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6262–6266. IEEE.

Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. 2020. Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews. *arXiv preprint arXiv:2006.08336*.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Linhai Zhang, Congzhi Zhang, and Deyu Zhou. 2023. Multi-relational probabilistic event representation learning via projected Gaussian embedding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6162–6174, Toronto, Canada. Association for Computational Linguistics.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. 2023. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*.

Deyu Zhou, Jianan Wang, Linhai Zhang, and Yulan He. 2021. Implicit sentiment analysis with event-centered text representation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6884–6893, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25(1):281–304.

A Appendix

A.1 Prompt Template

In this section, we present all the prompt templates employed for LLM-based methods in our experiments. The prompt for direct prompting is shown in Figure 4. The full prompt for personal query generation along with the basic query is shown in Figure 8. The Naive/Personal retrieval shares the same prompt, shown in Figure 5 as they are only different in the input queries. The preliminary assessment Prompt which is employed between the retrieval module and social intelligence enhancement module is shown in Figure 6. The event extraction prompt is shown in Figure 9. The full prompt for RED after the social intelligence enhancement module is shown in Figure 7.

A.2 Dataset Details

In this section, we provide an overview of the raw data included in the dataset, focusing solely on the Transcript. The corpus includes full textual transcripts of each interview, capturing both the interviewer's questions and the participant's responses. Detailed statistics can be found in Table 4.

Dataset	Size	Category	Round	Token
Train	107	[Deprssion] 30	6,069	149,149
		[Control] 77	$\bar{x} = 57$	$\bar{x} = 1,394$
Dev	35	[Deprssion] 12	1,909	53,588
		[Control] 23	$\bar{x} = 55$	$\bar{x} = 1,531$

Table 4: Detailed statistics of DAIC-WoZ.

The COKE benchmark is a cognitive knowledge graph for machine theory of mind (Wu et al., 2024). COKE contains a series of cognitive chains to describe human mental activities and behavioral/affective responses in social situations. It contains 5 dimensions, which are situation, thought, clue, action, and emotion. Detailed statistics can be found in Table 5.

Dimension	Count	Avg. Len.
Siutation	1,200	11.5
Thought	9,788	6.6
Clue	21,677	7.3
Action	19,875	6.8
Emotion	9,788	1.0

Table 5: Detailed statistics of COKE.

A.3 Implementation Details

Given the length of the clinical interviews, we have chosen a context window of 128,000 tokens. For our depression detection task, we currently use the following models: `gpt-4o-0806`, `gpt-4o-mini-2024-07-18`, and `gpt-4-0125-preview`.

In the evidence retrieval section, we divide the corpus into chunks of 500 tokens. Then, we use `text-embedding-3-large` embeddings to encode both the queries and transcripts, applying the L2 distance metric to retrieve the top K evidence (with$K = 10$ by default). In other experimental settings,K can also be 4, 6, or 8. For the social intelligence component, we consider two embedding methods. The first method is the same as the one used in evidence retrieval, while the second uses event triples, which are encoded using MORE-CL. Both methods retrieve the top M evidence (with$M = 2$ by default).

Direct Prompt
Below is a transcript of an interview between an interviewer and a participant. Based on the content of the interview, determine whether the participant is depressed or not. Attention: Answer 'DEPRESSION' or 'CONTROL' with explanations.

Figure 4: Prompt Template for Direct Prompting

Naive/Personal Prompt
Below are the key parts of the transcript from an interview between the interviewer and the participant. Based on the content of the interview, determine whether the participant is depressed or not by calculating the PHQ-8 scores. PHQ-8 Questionnaire Scores: 1. **Little interest or pleasure in doing things**: [0, 1, 2, 3] (How much did the participant report having little interest or pleasure in doing things?) 2. **Feeling down, depressed, or hopeless**: [0, 1, 2, 3] (How much did the participant report feeling down, depressed, or hopeless?) 3. **Trouble falling or staying asleep, or sleeping too much**: [0, 1, 2, 3] (How much did the participant report having trouble with sleep?) 4. **Feeling tired or having little energy**: [0, 1, 2, 3] (How much did the participant report feeling tired or having little energy?) 5. **Poor appetite or overeating**: [0, 1, 2, 3] (How much did the participant report having poor appetite or overeating?) 6. **Feeling bad about yourself — or that you are a failure or have let yourself or your family down**: [0, 1, 2, 3] (How much did the participant report feeling bad about themselves or like a failure?) 7. **Trouble concentrating on things, such as reading the newspaper or watching television**: [0, 1, 2, 3] (How much did the participant report having trouble concentrating?) 8. **Moving or speaking so slowly that other people could have noticed. Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual**: [0, 1, 2, 3] (How much did the participant report being slow or restless?) Calculate the total score based on the following rules: - **0** = No symptoms - **1** = Symptoms for a few days - **2** = Symptoms for more than half of the days - **3** = Symptoms nearly every day Add up the scores for all 8 items to get the total score. If the total score is greater than 9, determine that the participant is **DEPRESSION**. If the total score is 9 or below, determine that the participant is **CONTROL**.Answer with 'DEPRESSION' or 'CONTROL' and provide the explanation of the participant's PHQ-8 scores. **Output:** - **Total Score**: {Only total score based on your assessments} - **Final Diagnosis**: {CONTROL/DEPRESSION based on the total score} - **Explanation**: {Explanation of the participant's PHQ-8 scores and reasoning behind the final diagnosis.}

Figure 5: Prompt Template for Naive/Personal Prompt Retrieval

Preliminary assessment Prompt

You are a depression diagnosis expert. Below are evidences between an interviewer and a participant. Your task is to perform the following step:

Analyze the participant's symptoms and assess their severity based on the PHQ-8 principle.

For each of the following eight symptoms, provide a detailed analysis based on the evidences. Consider whether the symptom is present, its frequency, and how it affects the participant's well-being. Then, assess the severity of each symptom using the PHQ-8 scale (0 = Not at all, 1 = Several days, 2 = More than half the days, 3 = Nearly every day).

- **Little interest or pleasure in doing things**
- **Feeling down, depressed, or hopeless**
- **Trouble falling or staying asleep, or sleeping too much**
- **Feeling tired or having little energy**
- **Changes in appetite (poor appetite or overeating)**
- **Feeling bad about yourself, or that you are a failure, or have let yourself or your family down**
- **Trouble concentrating on things**
- **Moving or speaking slowly or being fidgety/restless**

Provide your reasoning and assessment for each symptom based on the evidences.

Output Format:

- **Little interest or pleasure in doing things**: {Your analysis}
- **Feeling down, depressed, or hopeless**: {Your analysis}
- **Trouble falling or staying asleep, or sleeping too much**: {Your analysis}
- **Feeling tired or having little energy**: {Your analysis}
- **Changes in appetite**: {Your analysis}
- **Feeling bad about yourself**: {Your analysis}
- **Trouble concentrating**: {Your analysis}
- **Moving or speaking slowly or being fidgety/restless**: {Your analysis}

Guidelines:

- Be sure to analyze each symptom based on the participant's statements in the evidences.
- Use the PHQ-8 criteria to rate the severity of each symptom.
- Ensure your reasoning is thorough and consistent with the principles of depression diagnosis.

Figure 6: Prompt Template for Preliminary Assessment

RED Prompt

You are a depression diagnosis expert. Below is the preliminary symptom assessment result along with recent major events and ToM (Theory of Mind) samples that provide insights into the participant's perception of others' mental states. Your task is to **re-evaluate** the severity of the participant's symptoms by deeply analyzing the major events, selectively integrating relevant ToM insights, and considering the broader psychological impact of these experiences.

Please follow the steps below to ensure a thorough reassessment:

Step 1: Contextual Analysis of Major Events and Selective ToM Insights

Start by analyzing each **major life event** individually, thinking expansively about its potential psychological impact. Consider:

- How might this event **directly or indirectly** contribute to depressive symptoms?
- What **emotional, cognitive, and behavioral** responses might typically arise from this experience?
- Are there **secondary effects** (e.g., social withdrawal, self-doubt, altered self-perception) that could reinforce or trigger additional symptoms?

Next, **selectively integrate relevant ToM insights** by considering:

- Does the participant's perception of others' emotions, thoughts, or intentions **amplify or mitigate** the event's impact?
- Do ToM biases (e.g., excessive guilt, misinterpretation of others' behavior, heightened social comparison) **exacerbate depressive symptoms**?
- Could the participant's social and emotional interpretations **shape the way they process these events**, either positively or negatively?

Summarize the **most relevant insights**, ensuring that they meaningfully contribute to symptom reassessment.

Step 1 Output:

-**Expanded Psychological Analysis of Major Events**: {For each major event, explore its direct and indirect effects on emotions, thoughts, and behavior.}

-**Selective ToM Insights and Their Influence**: {Summarize only the most relevant ToM insights and explain how they interact with major events.}

Step 2: Holistic Symptom Reassessment

Now, use the **PHQ-8 scale** to reassess each symptom, integrating the expanded event analysis and selective ToM insights. Reflect on:

- How does **each major event** contribute to this symptom, either as a trigger or reinforcing factor?
- Does the participant's **social-cognitive interpretation** of these events **intensify or alleviate** this symptom?
- Are there **unconscious patterns or secondary effects** that might explain symptom persistence or fluctuation?

PHQ-8 Symptom Scale: - **0 = Not at all**- **1 = Several days** - **2 = More than half the days** - **3 = Nearly every day**

Step 2 Output:

- **Little interest or pleasure in doing things**: {Updated score and explanation, considering major events and ToM influences.}
- **Feeling down, depressed, or hopeless**: {Updated score and explanation, considering major events and ToM influences.}
- **Trouble falling or staying asleep, or sleeping too much**: {Updated score and explanation, considering major events and ToM influences.}
- **Feeling tired or having little energy**: {Updated score and explanation, considering major events and ToM influences.}
- **Changes in appetite**: {Updated score and explanation, considering major events and ToM influences.}
- **Feeling bad about yourself**: {Updated score and explanation, considering major events and ToM influences.}
- **Trouble concentrating**: {Updated score and explanation, considering major events and ToM influences.}
- **Moving or speaking slowly or being fidgety/restless**: {Updated score and explanation, considering major events and ToM influences.}

Step 3: Final Score and Diagnosis

Using the updated symptom ratings from **Step 2**, calculate the **total PHQ-8 score** and determine the **final diagnosis**.

PHQ-8 Scoring Reference:- If the total score is **0–9, the participant is classified as **Control**.- If the total score is **10 or higher**, the participant is classified as **Depression**.

Step 3 Output:

- **Total Score**: {Final total score based on reassessment}
- **Final Diagnosis**: {CONTROL/DEPRESSION based on the total score}

Figure 7: Prompt Template for RED Final Assessment

Personal information generation Prompt

You are a helpful assistant. Below is a transcript of an interview between an interviewer and a participant. Based on the transcript, please summarize the basic information of the participant. Provide only the summary, without any additional explanations. The summary should include key details such as occupation, mood, or any relevant personal or emotional information mentioned by the participant.

Personal query generation Prompt

You are now a professional psychologist. Based on the following basic information about a participant, please provide personalized consultation. Ask questions specifically related to the eight aspects from the PHQ-8 questionnaire. The eight aspects are:

1. Little interest or pleasure in doing things
2. Feeling down, depressed, irritable, or hopeless
3. Trouble falling or staying asleep, or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself, or that you are a failure, or have let yourself or your family down
7. Trouble concentrating on things (e.g., school work, reading, watching television)
8. Moving or speaking slowly that others may have noticed, or feeling fidgety or restless

Please formulate personal questions related to these aspects based on the information provided by the participants. Present your question in a list format without any additional detailed information required.

Iterative query generation Prompt

You are tasked with generating a new query based on the provided `existing query` and `evidences` from a conversation. The goal is to generate a follow-up question that delves deeper into the same topic, maintaining consistency with the current focus (e.g., symptoms related to PHQ-8 or mental health issues) and providing an opportunity for more detailed responses.

1. **Existing Query**: {existing_query}
2. **Evidences**: {evidences}

Generate a new query that focuses on the same topic as the existing query but requests additional details or clarifications, with the goal of further exploring the participant's experiences, symptoms, or thoughts related to the topic. The new query should not introduce a different topic or symptom area, but instead should build upon the existing information.

Guidelines:

- The new query should be a direct follow-up to the existing query.
- It should be relevant to the existing context, e.g., if the existing query is about sleep disturbances (a PHQ-8 symptom), the new query should be related to sleep or similar topics.
- Do not deviate from the current topic. If the existing query is about depression symptoms, the follow-up should remain on that topic.
- The new query should aim to gain more insight into the topic or clarify existing responses.

Example Input:

- Existing Query: "Have you been feeling more down or hopeless lately?"
- Evidences: ["Participant: I've been feeling really low the past few weeks.", "Ellie: Can you describe more about how you're feeling?"]

Output:

- "Can you tell me more about what activities or situations make you feel particularly hopeless or down?"

Evidence check Prompt

You are given the following `evidences` from a conversation with a participant. Your task is to determine whether these evidences are sufficient to assess whether a specific symptom or condition (e.g., related to PHQ-8 or mental health issues) is present or not.

1. **Symptom**: {symptom}
2. **Evidences**: {evidences}

Based on the provided evidences, do you believe you have enough information to make an assessment of this specific symptom?

Please respond with:

- **Yes** if the evidences are sufficient to make a judgment on the symptom.
- **No** if the evidences are insufficient, and more information or further questions are needed to make a clear judgment.

Example Input:

- Evidences: ["Participant: I haven't been able to sleep for the past week.", "Ellie: How often do you have trouble falling asleep?"

Output:

- No

Figure 8: Prompt Template for Query Generation

Event extraction (transcript) Prompt

Extract all major events related to the participant from the following conversation and represent each event as a triplet in the format: <subject, predicate, object>.

- Subject: The person performing the action (e.g., Participant).
- Predicate: The action or verb (e.g., ask, respond, go).
- Object: The recipient or target of the action (e.g., park, plans, person).

Example:

Sentence: 'You abandon me for a week to go off on holiday with daddy, come back and barely 2 days later you go off out with him again.'

Event triplet: <you, abandon, me>

For each event, ensure the relationship is clear and follows the structure <subject, predicate, object>. If the event is unclear or you cannot extract a meaningful triplet, skip it and output "No".

Output Format:

Triplets in the format <subject, predicate, object> and Separate triplets with '\n'.

Event extraction (coke) Prompt

Please extract key event from the following information and represent it as event triplets in the format: <subject, predicate, object>. Each triplet should reflect a clear subject-action-object relationship. Subject: The person performing the action (e.g.,I). Predicate: The main verb that describes what the subject is doing (e.g., ask, respond, go, discuss). Object: The recipient or the thing that the action is directed towards (e.g., park, plans for the weekend). Example: Sentence: 'You abandon me for a week to go off on holiday with daddy, come back and barely 2 days later you go off out with him again.' <you, abandon, me>. Finally, show your answer in the format: <subject, predicate, object> of a list. If the event is not clear or no event can be extracted, do not include it."

Figure 9: Prompt Template for Event Extraction