Detection of Depression with Symptom Similarity: Data Reduction and LLM Personas

Notebook for the eRisk Lab at CLEF 2025

Ane Varela[1,2,*,†], Maite Oronoz[1,2], Arantza Casillas[1,3] and Alicia Pérez[1,2]

[1]HiTZ Basque Center for Language Technology (hitz.eus)- Ixa, University of the Basque Country UPV/EHU, Spain
[2]Department of Computer Languages and Systems - UPV/EHU
[3]Department of Electricity and Electronics - UPV/EHU

Abstract

This paper presents the approaches presented on the eRisk 2025 initiative, aimed at the early detection of mental health risks through the analysis of users' social media. On the one hand, we explore the first task, in which a user ranking has to be obtained to seize relevance of depression symptoms in user's writings. The symptoms are based on the second revision of the Beck Depression Inventory questionnaire, BDI-II. We deal with a transformer-based approach for ranking using a multilingual BERT model with multiple heads, and a two novel data reduction techniques to optimize training and inference time. Our results show the effectiveness of these selection strategies, as reduced inference time was obtained while achieving higher performance. On the other hand, we explore the Pilot Task. This consists of developing conversational agents that can interact with Large Language Model (LLM) personas and detect whether the latter have depression, also based on depression symptoms. We employed various LLMs for this task, including GPT and Falcon, evaluating their conversational assessments of depression. Our results indicate that LLMs, even without fine-tuning, can perform comparably to manual approaches in symptom identification and severity estimation, with GPT-based systems showing the most promise in balancing brevity and informativeness.

Keywords

semantic similarity, data reduction, depression symptoms, LLM personas

1. Introduction

Even though the awareness for mental health issues has increased in the last years, there is still stigma related to mental illnesses and their treatment [1]. This is a concern, as symptom management of these issues is influenced by the mental health literacy, that is, the knowledge and acceptance around mental health of the person with the sickness or those around them. Taking into account that almost all people will come in contact with someone impaired by a mental illness [2], it is crucial to improve public awareness and knowledge regarding this topic. In general, it is clear that some work needs to be done in order to better integrate patient needs in this particular healthcare field.

The widespread use of social media presents a unique opportunity in this regard. Social platforms generate large volumes of user-generated content that can be analysed to identify early signals of mental health deterioration, including depression and suicidal ideation [3]. Moreover, the growing volume of online research studies, even if it supports scientific advancement and promotes information sharing, also presents challenges for mental health professionals who must navigate and synthesize an overwhelming amount of data, which can lead to stress and burnout and even reduce decision-making abilities [4]. Therefore, methods based on informatics, and, nowadays, Artificial Intelligence (AI) have arisen as an opportunity to reduce the administrative burden on these workers, and to act as a support to access information, among others [5].

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
ane.varela@ehu.eus (A. Varela)
0009-0000-9426-9681 (A. Varela); 0000-0001-9097-6047 (M. Oronoz); 0000-0003-4248-8182 (A. Casillas); 0000-0003-2638-9598 (A. Pérez)

In this context, the *eRisk* shared task, organized within the CLEF (Conference and Labs of the Evaluation Forum) workshop, provides a standardized benchmark for the early risk detection of mental health issues using social media data [6, 7]. The eRisk shared task is one of the initiatives in which current advances in informatics can be applied to the mental healthcare field. Our team participated in two of the subtasks:

- **Task 1: Early Risk Detection of Depression**. This task involves ranking user-written sentences according to their relevance to depression symptoms according to the BDI-II questionnaire. This ranking is evaluated by comparing it to human annotation, which has two possibilities: label assignment by majority and by unanimity.
- **Pilot Task: Conversational Depression Detection via LLMs**. This task, new for this edition, introduces a challenge of interacting with Large Language Model (LLM) personas. The challenge lies in determining the signs and symptoms of depression of the persona, and determining their possible score in the BDI-II questionnaire.

The main NLP-related challenges of the eRisk tasks include:

1. Detection of implicit expressions of mental states, and, particularly, depression symptoms.
2. Dealing with imbalanced and noisy data, including instances in languages other than English.
3. In Task 1 specifically, consensus versus majority label disagreements add another layer of complexity, where reliability of annotations varies across examples.

Our group, **ixa_ave**, explored methods informed by prior work in the eRisk framework, tailoring them to the current task and challenges with our own ideas. The following section reviews the current landscape of NLP in mental health detection and in the eRisk tasks, and lays the foundation for our methodology.

2. Related work

Historically, the way to assess a person's mental state has been through a thorough psychological analysis performed by mental healthcare experts. However, as the number of people needing access to this kind of service increases, rule-based (RB) methods based on questionnaire assessment arose to be able to assess patients individually in a quicker manner. These methods normally consist of sets of questions that infer the most relevant aspects for a certain mental health issue, including the detection of suicidal intent [8]. This method, although more efficient, is not without its limitations, mostly regarding the potential of patients to lie more easily and a reliance on predefined questions that may not capture the complexity of a patient's experiences [9]. However, as large amounts of textual data from social media have become available for training models, the use of Natural Language Processing (NLP) has been proposed as an alternative to determine a person's mental state.

For instance, simple Machine Learning (ML) techniques such as logistic regression, Support Vector Machines (SVMs), and decision trees have been used successfully in various tasks of detecting mental health conditions. These methods usually outperform rule-based approaches by capturing more complex patterns in the data, such as non-explicit expressions of suicidal ideation, or changes in tone [10].

On the other hand, more complex machine learning approaches, like neural networks, provide a tailorability that can be effectively used for mental health applications. Due to the ability of models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture both short-term and long-term dependencies in text, deep learning models have significantly contributed to NLP [11, 12]. The use of dense vectors to represent sentences also give these models the ability to better capture semantic similarities [13]. All of these benefits make the deep learning approach more adaptable and effective in NLP than traditional ML approaches. Furthermore, word and sentence similarity metrics have become increasingly valuable in assessing mental health issues by analysing the semantic content of responses [14].

However, these type of architectures struggle with long-term dependencies. Arising to this need, pre-trained transformer models such as the Bidirectional Encoder Representations from Transformers (BERT) have transformed the field [15]. Transformers excel in transfer learning, where pre-trained models can be fine-tuned for specific tasks. These pre-trained models already have semantic knowledge embedded in them, and thus fine-tuning becomes much less computationally expensive and time-consuming, while allowing the model to rapidly specialize to domain-specific tasks with minimal data, and delivering strong performance gains. This approach significantly reduces the need for large labelled datasets, which are often expensive and time-consuming to obtain. For instance, for the specific processing of medical texts, some specific transformers have been trained, like MedicalBERT or BEHRT [16]. These can be used in the specific case of mental health with good results [17].

Some more recent examples of applications of NLP have been Large Language Models, or LLMs. These have been successfully applied to the healthcare department, for instance, for differential diagnosis [18] and medical text summarization [19], and also in more mental health related cases, like for suicide assessment [20]. These recent studies, along with the boom of LLMs, demonstrate that there is a great potential in LLMs for this kind of health applications.

Over the past editions of the eRisk shared task, several research groups have tackled the first task, exploring a wide range of strategies. Traditional machine learning pipelines initially dominated the field, with participants performing feature engineering to retrieve the most significant lexical cues [21]. With the rise of deep learning, more recent approaches have adopted models like RNNs to capture temporal and semantic dynamics across user timelines [22]. However, most teams use transformer-based architectures, such as BERT and its derivatives, with excellent results [23, 24]. Additionally, methods like curriculum learning, weak supervision, and meta-learning have been proposed to handle the sparse and imbalanced nature of user-level labels. More recent approaches have also focused on retrieving or synthesizing data with LLMs to improve depression detection. For example, [25] proposed using ChatGPT-generated data to retrieve depression symptoms from social media, thus improving the performance. However, most prior approaches do not directly address the semantic alignment between training examples and clinical symptom definitions.

In order to take this semantic alignment into account, our idea focuses on a similarity-based approach, where the semantic similarity between the assessed sentences and the BDI-II questionnaire items was taken into account. Prior work has approached this problem through filtering systems to reduce the search space [26], as well as similarity mechanisms calculated with respect to the questionnaires [27]. While these methods have laid important groundwork, they often rely on rigid similarity metrics or static thresholds that may not adapt well to the subtle and context-dependent nature of social media posts. Furthermore, they lack a fine-grained semantic alignment with established clinical frameworks, such as the BDI-II questionnaire. To address these limitations, our work introduces two data reduction strategies based on semantic similarity to the BDI-II. This approach improves the relevance of selected content, optimizes training and inference time, and enhances the interpretability of the classification process, distinguishing our methodology from prior efforts in the eRisk series and contributing a new, clinically informed dimension to risk detection.

In the Pilot Task, as it was its inaugural year, we faced the challenge of charting new territory, as no directly comparable prior work existed. Nevertheless, we were inspired by existing applications of LLMs in diagnostic contexts [18] and their demonstrated capacity to assess personality traits [28]. Therefore, we hypothesized that leveraging LLMs as evaluators themselves could give valuable insights. Our goal was to comparatively investigate how both open-source and proprietary LLMs would perform relative to humans, especially in the absence of any specialized mental health training. To that end, our methodology was intentionally minimalistic: the developed models were provided only with the BDI-II questionnaire and stylistic prompting before interacting with the simulated personas they were tasked with assessing.

3. Materials and methods

3.1. Materials

For the Pilot Task, no training data was made available. By contrast, for Task 1, a TREC-formatted sentence-tagged dataset was provided. The dataset consisted of user-generated Reddit sentences, with each sentence associated with one or more depression-related symptoms from the BDI-II questionnaire (see the first column in Table 1) as annotated by experts. Each user-generated sentence included metadata following an XML-like structure:

- <DOCNO>: A unique identifier for the sentence, used for referencing.
- <PRE>: The sentence that comes immediately before the target sentence in the user's post or timeline, in order to provide context.
- <TEXT>: The target sentence that is asked to evaluate.
- <POST>: The sentence that comes immediately after the target sentence, offering additional context.

Not all training data had the <PRE> and <POST> data, so we chose to not use them for simplicity's sake. Apart from this, for training purposes, two comma separated value `csv` files were released. These files identified sentences from the training dataset (thanks to the document identifier, <DOCNO>) that were relevant to specific BDI-II symptoms, based on human annotation:

- **Majority** vote. This `csv` included sentences that were labelled as relevant by the majority of annotators. This reflects a more inclusive judgment criterion and is useful for training models with greater coverage.
- **Consensus** vote. This `csv` contained only those sentences that received unanimous agreement or full agreement among annotators about their relevance. These are considered high-confidence labels.

Not all the training data was annotated; in fact, more than 99% of the training dataset was unlabelled. More particularly, a sentence could be annotated for a symptom but not the rest of the symptoms. The dataset was also imbalanced, with more sentences being classified as irrelevant (0) than as relevant (1). The amount of data for each symptom can be seen in Table 1.

Furthermore, a main characteristic of our work has been the use of the BDI-II questionnaire [29] for both tasks. In the first task, it has been used to calculate the messages' similarity with respect to the depression symptoms, and in the Pilot Task, it has been used as a prompting agent for the developed LLM personas. The left column in Table 1 shows the 21 symptoms related to depression according to this questionnaire.

3.2. Methods

3.2.1. Task 1: Search for Symptoms of Depression

For the depression symptom ranking task, we implemented several models based on the original multilingual BERT architecture, that is, the *bert-base-multilingual-cased*[1] model [15]. All the models were fine-tuned on an Nvidia A100 GPU on two epochs with an `AdamW` optimizer and a learning rate of $2e-5$. The batch size was 32 and the maximum token length was 512. For reproducibility purposes, the code for this Task has been uploaded to https://github.com/anevarela/eRisk_ixa_ave.

The models tackled the task as a classification, and they had a 21-head output with a `softmax` activation function, enabling an inference for each sentence to all symptoms in the BDI-II questionnaire in a regression-like manner. Moreover, all of the models introduced the similarity of each sentence with respect to the 21 symptoms of the BDI-II questionnaire as training features. The output of the employed

[1]Available at https://huggingface.co/google-bert/bert-base-multilingual-cased

Table 1
Label counts per symptom for Majority and Consensus categories.

Symptom	Majority		Consensus	
	0	1	0	1
1. Sadness	1020	641	1313	348
2. Pessimism	1077	540	1427	190
3. Past failure	976	509	1199	286
4. Loss of pleasure	1100	363	1274	189
5. Guilty feelings	888	460	1010	338
6. Punishment feelings	1294	178	1380	92
7. Self-dislike	945	589	1131	403
8. Self-criticalness	1137	408	1305	240
9. Suicidal thoughts or wishes	792	631	941	482
10. Crying	847	675	1036	486
11. Agitation	1103	473	1281	295
12. Loss of interest	1186	337	1342	181
13. Indecisiveness	1212	355	1391	176
14. Worthlessness	1113	350	1216	247
15. Loss of energy	1098	405	1224	279
16. Changes in sleeping pattern	863	631	1103	391
17. Irritability	1142	377	1281	238
18. Changes in appetite	973	492	1181	284
19. Concentration difficulty	1045	378	1165	258
20. Tiredness or fatigue	1018	486	1193	311
21. Loss of interest in sex	1153	382	1335	200

multilingual BERT was concatenated with the 21 similarity values and fed to a final layer, that was then connected to the classifier heads.

The **similarity** of the sentences with respect to the 21 BDI-II questionnaire items was computed by means of Cosine Similarity. The model employed to transform the text into numerical representation was the Sentence Transformer *mrm8488/distiluse-base-multilingual-cased-v2-finetuned-stsb_multi_mt-es*[2]. The similarity with respect to each symptom was calculated using the questionnaire's items. Each symptom of the BDI-II has a severity ranging from 0 to 3 in a Likert scale point system, based on different statements. Take the following example from the BDI-II questionnaire:

> **14. Worthlessness**
> 0. I do not feel I am worthless.
> 1. I don't consider myself as worthwhile and useful as I used to.
> 2. I feel more worthless as compared to others.
> 3. I feel utterly worthless.

This shows that each symptom can have different degrees of severity, 0 (less severe) to 3 (more severe). To compute the similarity of one sentence of the dataset with respect to one of the 21 symptoms, we apply a weighted similarity across the symptom's severity levels. The assigned weights, w_j range from 1.0, for severity 0, to 2.5, for severity 3. This reflects the increasing diagnostic value and specificity of more severe statements, as the weights increased linearly with respect to severity with a step of 0.5.

While low-severity descriptions still indicate the user's condition and are relevant, they often express absence of the symptom and tend to be more general in terms of formulation. In contrast, high-severity descriptions carry more distinct signals of symptom presence, making them more informative for relevance estimation. Normalizing by the sum of weights ensures comparability across symptoms and bounds the final score, as it can be seen in expression (1), which states the computation of the similarity

[2]Available at https://huggingface.co/mrm8488/distiluse-base-multilingual-cased-v2-finetuned-stsb_multi_mt-es

for each item of the questionnaire, Q_i, related to the i-th symptom.

$$wSim(S, Q_i) = \frac{\sum_j w_j cos(S, Q_{ij})}{\sum_j w_j} \tag{1}$$

In this expression, the similarity of a sentence S with respect to the questionnaire item Q_i related to a certain symptom is calculated. The calculation is normalized with the sum of all weights.

Regarding the use of the data, a 20% of the provided training set was used for the validation, randomly selecting users for this purpose. Moreover, oversampling was also introduced in the system to avoid classifying to the most prominent class, 0, as the data was highly imbalanced. It is also important to note that not all the training and validation instances were ultimately used, as only the labelled data were used for this (see Table 1). This is because, even if, ideally, there would be a pre-training stage to get the model used to the language of the desired task [30], it was skipped in this approach due to time and memory constraints. If the label of a certain symptom was not known, a -1 mask was used to mark it. These masked sentences were not used to compute the loss at the pertinent symptoms.

Additionally, a weighting mechanism was incorporated into the loss computation to distinguish between consensus and majority annotations. Specifically, predictions with consensus labels were given twice the weight of those with only majority labels, emphasizing high-agreement samples during training, as it can be seen in expression (2). Here, ℓ_{BCE} denotes the binary cross-entropy loss, p_i is the probability of the predicted label, and y_i is the true label. Note that, whenever the y_i annotation was made by consensus, i.e. $consensus(y_i) = 1$, thus, the weight results in 2α, that is, twice as much as with mere majority annotation.In practice, we selected $\alpha = 1$.

$$\ell_{\text{weighted}} = \alpha(1 + \delta(consensus(y_i))) \cdot \ell_{\text{BCE}}(p_i, y_i) \tag{2}$$

The specifications above are common for all the models.

Although it may seem counterintuitive, instead of regular data augmentation approaches, we turned to data reduction. Formally, our approach focuses on a subset of significant sentences with respect to symptoms, both in the training and test stages. The training set data is reduced in an attempt to keep sentences that are not really obvious, as in (3); only those sentences with similarities of less than a threshold β are kept for training. Regarding the test sentences used to make the decision, only those that are relatively connected to symptoms are employed, as in (4). Only those sentences with similarities higher than a threshold θ are kept for inference.

$$Train_\beta = \{t = (X, Y) \in Train \mid \exists Q_i : wSim(X, Q_i) \leq \beta\} \tag{3}$$

$$Test_\theta = \{t = (X, Y) \in Test \mid \exists Q_i : wSim(X, Q_i) \geq \theta\} \tag{4}$$

As a result, our approach is characterized by two parameters (β, θ). Based on this, our team submitted **five runs**:

1. **Base model**. This model was trained in all the labelled training data using gradient descent. We submitted three runs based on this:
   a) **`base_all`**. The inference was conducted in all the provided test data.
   b) `base_filter`. This strategy included what we will call the **test data selection** strategy. The approach consisted of ignoring all the instances from the test set that had a similarity, as in (1), **below** a certain θ. The rationale behind this is that they were deemed semantically unrelated or off-topic, and eliminating them before feeding them to the model would make the inference process much quicker; the greater the value of θ, the fewer instances will be given to the model. Two different values of θ were explored, 0.3 and 0.5, leading to the so-called **`base_filter30`** and **`base_filter50`**, respectively.
2. **Threshold model**. This model introduced a **training data reduction**, under the assumption that sentences with a high similarity **above** a threshold β would yield little additional learning

value (i.e. they would be "too easy" to learn) and, thus, could be effectively skipped during training, as their associated loss would likely approach zero. A $\beta = 0.5$ was selected. Two runs were submitted with this training strategy:

a) **`thresh_all`**. Only the training data reduction was applied, with $\beta = 0.5$. Thus, the inference was conducted in all the provided test data.
b) **`thresh_filter50`**. Apart from the training data reduction with $\beta = 0.5$, the test data selection strategy was applied, with a $\theta = 0.5$, ignoring all sentences from the test dataset having a similarity with respect to the symptoms below this θ.

A small summary of the models can be seen in Table 2. Notice that, for $\beta = 1$, no sentence is going to have a higher similarity, and thus all the training data is taken into account; similarly, for $\theta = 0$, no test data is eliminated, as the similarity will always be higher.

Table 2
Run specifications: the models were trained and tested on different sub-set of the original training set, based on of similarity of the sentences with respect to the questionnaire. The training subset was selected based on a threshold β, with sentences higher than it being excluded. The test dataset excluded sentences lower than a threshold θ.

run	β	θ
`base_all`	1	0
`base_filter30`	1	0.3
`base_filter50`	1	0.5
`thresh_all`	0.5	0
`thresh_filter50`	0.5	0.5

Our work explores a combination of *training data reduction* (by excluding obvious training samples with β) and *test data selection* (by removing unrelated samples with θ). This was validated in preliminary experiments and is consistent with the obtained results.

3.2.2. Pilot Task: Conversational Depression Detection via LLMs

The Pilot Task for this year consisted of being able to develop a chatbot that would interact with twelve proposed personas. These personas would each have a certain depression score and symptoms associated to them, and the objective was to automatically infer these scores and symptoms.

We explored three primary approaches to conducting depression screening conversations. All developed approaches can be described as minimally trained base models (or personas) assigned the task of assessment based on the BDI-II questionnaire. That is, none of them were explicitly trained or fine-tuned in a dataset related to mental health or depression. In contrast, the systems were prompted with the questionnaire and the task in hand, with no further information. This approach was mimicked in the manual approach, in which a human with no specific mental healthcare knowledge conducted the conversation, once given the BDI-II questionnaire.

Four different runs were presented. One run was manual, and the other three were based on LLMs. Two of the LLM approaches were based on the GPT-4 model [31], and the other one was based on an open-source model, Falcon [32]. This breakdown of all approaches explains each one with more details, and specifies the run related to them:

- **Manual approach** (run 0). A human without specific clinical training conducts the conversation using the BDI-II questionnaire as a base. The dialogue is then analysed using a model previously trained for Task 1 to infer a BDI-II score and identify key symptoms, if any. The used model was the `base_all`.
- **GPT-based agents.** Two versions of GPT personas were implemented:
  - *Long GPT*[3] (run 1). This version systematically goes through all BDI-II symptoms in detail and in order. The initial prompt to determine the system instructions emphasized that it was

[3]Available from https://chatgpt.com/g/g-67e55a078fd4819180b2a3d3651dd7e3-supportive-mood-checker

a supportive persona in a therapist-like style, that needed to analyse depression symptoms as stated in the questionnaire in a conversational and emphatic manner, and then infer a score based on the used questionnaire. In the conversations, it was prompted to start with a general question and then, once the conversation was concluded, it gave a BDI-II score and top symptoms in the desired format. The BDI-II questionnaire was given as a reference document. The score for each symptom is not calculated; instead, the whole conversation is assessed and given an overall depression score.

  - *Short GPT*[4] (run 2). This version's system instructions included a brevity aspect, asking so that it combined symptoms into fewer prompts to complete the assessment more quickly. This was based on the idea that an earlier detection would be beneficial, as it was stated in the task definition that a longer assessment would result on penalisation. Like its longer counterpart, this approach was prompted to start with a general question and ended with the overall inference of the persona's mental state based on the BDI-II questionnaire. The BDI-II questionnaire was given as a reference document.

- **Open-source LLM (Falcon)** (run 3). We employed the Falcon model, a lighter alternative to LLaMA [33], which, when given the task description, generates one synthetic question per BDI-II symptom, with a maximum length of 50 tokens. A second instance of the model then infers a BDI-II score and highlights key symptoms based on the responses. This approach also relied on the use of the BDI-II questionnaire scoring system and calculated the score for each symptom individually. The overall score was taken as a combination of all 21 symptoms, like it is done in the BDI-II questionnaire.

All the prompts for the used LLM models (both GPT and Falcon) can be found in Appendix A.

4. Results

4.1. Task 1: Depression ranking task

The models developed for this task, after inference of the test set, were evaluated using widely used information retrieval metrics [34]. **Mean Average Precision (AP)**, **Mean R-Precision**, **Mean Precision at 10 (P@10)** and **Mean NDCG at 1000** were used. These assess the effectiveness of the models in retrieving and ranking relevant items.

In the preliminary results obtained with the validation partition, applying **test data selection** appears viable for the Base model, but less so for the Threshold model, as this also included **training data reduction**. This distinction is intuitive: the Threshold model was never exposed to high-similarity cases during training, and thus lacked the capacity to generalize effectively to such instances at inference time.

Similarly, if this task were taken as a classification instead of a ranking based on the typical 0.5 limit, the validation metrics seem to confirm that the Base model performs better than the Threshold model. A F1-score of 0.85 is obtained with the Base model with no test data selection, while a 0.81 score is obtained with the Threshold model under the same test conditions. Nevertheless, employing similarity-based training data reduction remains promising, as it simplifies the learning process and reduces computational demands. In our case, 3.65% of the training labels were skipped due to this training data reduction, leading to a slightly more efficient approach. An efficient trade-off may be found, but this strategy exhibited limitations and our assumption that the high-similarity sentences may be "too obvious" may be incorrect.

The two `base_filter` runs show that applying **test data selection** can be beneficial. In fact, it can be confirmed during validation that the number of False Positives (FP) is reduced, at the cost of increasing the number of False Negatives (FN), when test data selection is applied. Figure 1 visualizes this degradation using a Sankey diagram, which illustrates the evolving confusion matrix as the test data selection increases.

[4]Available from https://chatgpt.com/g/g-67ee482cf2dc819195533b897fd63630-depression-screening-assistant

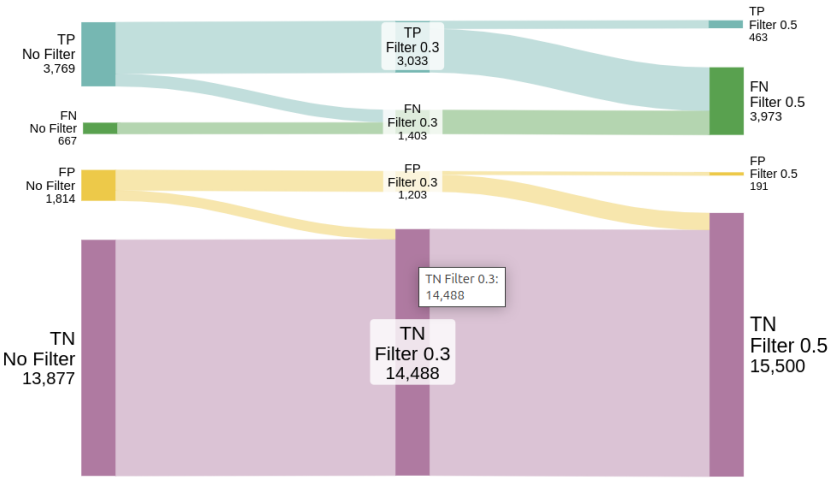

Figure 1: Sankey diagram showing the change in the confusion matrix of the validation set depending on the used filter for the Base model.

However, and as expected, the test data selection does contribute to the acceleration of the inference process. When ranking the test partition, for the 0.3 filter, more than 94% of instances were excluded because of low similarity with respect to the questionnaire, causing the inference time to drop to only a 5% of the time needed for all the instances. This seems to improve performance of the Base model by dropping a number of false positives, as it can be seen in Tables 3 and 4, which show our team's performance in majority and unanimity voting, respectively. The difference is even greater with the 0.5 filter; although the performance is poor, the excluded sentences surpass 99%. This suggests a compromise can be found based on the trade-off between time and performance.

Table 3
Performance of different system configurations on Task 1 (majority voting).

Run	AP	R-PREC	P@10	NDCG
base_all	0.097	0.191	0.305	0.345
base_filter30	0.102	0.203	0.338	0.342
base_filter50	0.009	0.025	0.086	0.048
thresh_all	0.091	0.168	0.281	0.333
thresh_filter50	0.005	0.016	0.129	0.035

Table 4
Performance of different system configurations on Task 1 (unanimity voting).

Run	AP	R-PREC	P@10	NDCG
base_all	0.053	0.121	0.124	0.282
base_filter30	0.055	0.126	0.138	0.277
base_filter50	0.006	0.020	0.038	0.042
thresh_all	0.052	0.103	0.110	0.270
thresh_filter50	0.003	0.013	0.048	0.026

Our AP levels are not high, and they suggest that out of 10 retrieved sentences, only 1 or less is considered relevant across the full list. In contrast, our P@10 of 0.30 means 3 of the first 10 results are relevant. R-PREC evaluates precision at the rank equal to the number of true relevant items, and only achieves a maximum of 0.305 in our models. Finally, NDCG accounts for the position of relevant items in the ranked list, and only rises up to 0.345 for us.

Our best-performing model was base_filter30, achieving 0.102 AP in majority voting and 0.055 AP in unanimity. Comparing with the rest of the groups, our team's performance was mediocre, particularly

on the majority voting [6]. It is interesting to see that, comparing to the information loss some groups have from majority voting, our model seems to be more robust to this; our AP drop between majority and unanimity was only 0.047, whereas the average drop across top 5 teams was 0.1. Therefore, we have a strong conviction that our approach may be useful with a better fine-tuning, and particularly, that our approach to include weighted loss in the consensually labelled sentences is appropriate.

4.2. Pilot Task: Depression chatbot

To value the performance of the Pilot Task, three main evaluation metrics were used, adapted from eRisk 2019 [35]: **Depression Category Hit Rate (DCHR)**, **Average DODL (ADODL)**, and **Average Symptom Hit Rate (ASHR)**. The metrics shown in Table 5 are in accordance with the described models. The mean number of messages per run define the number of prompts generated by our models (or written by the user) to send to the personas.

Table 5
Generated message characteristics based on run. (*) specifies a manual run.

#Run	#Mean messages per run	#Mean characters per message (by proposed model)	#Mean characters per message (response by persona)
0*	9.08	58.91	287.2
1	21.58	208.80	579.20
2	8.67	594.86	1009.07
3	22.67	95.41	669.99

As it can be seen in Table 5, run 0, related to the manual approach, and run 2, the short GPT model, are the ones that send the least number of messages, and while the short GPT model does pack a lot of information in the messages (both in the questions and the responses to the personas), the manual approach is poorer; the prompts are short, and the responses also seem to lack information based on the number of characters. On the other side, run 1 and 3, which are based in the long GPT approach and Falcon approach respectively, opt for a longer interview-style prompting, with approximately one question per symptom. This seems to work for the GPT model, as it develops richer questions and therefore obtains long answers, but the Falcon model, although more concise in its questioning, also obtains long answers from the personas.

The longest responses from the personas correspond to the short GPT model, maybe due to the information packing that comes from asking about several symptoms at the same time. This is also the model with the least number of prompts, making it the most appropriate for quick assessment.

The metrics obtained by our group in this shared task can be seen in Table 6. The asterisk (*) represents the manual run; it is important to mention that ours was the only manual run submitted.

Table 6
Pilot Task results for our team, ixa_ave.

Run	DCHR	ADODL	ASHR
0*	0.33	0.80	0.25
1	0.33	0.76	0.29
2	0.33	0.83	0.21
3	0.17	0.81	0.19

Out of the four presented models, the results given by the evaluation metrics favour the GPT models. The manual and GPT-based models all score a DCHR metric of 0.33, showcasing limited capability of the models to effectively predict the depression category of the personas. A random guess in the DCHR metric has a 25% chance of hitting the correct category; our slightly higher metric showcases an increase in predictive capacity in this matter, with 4 out of 12 personas being classified correctly. However, the Falcon LLM model, showcasing a DCHR of 0.17, shows less capability than a random model for this severity classification.

The DCHR scores seem to be in accordance with the expected results, as the distribution of inferred BDI-II scores does not change significantly across models, except in the case of the Falcon model, which produces significantly lower results in general. This is shown in Figure 2. This is also corroborated by a statistical significance test, that showcases no differences between Runs 0, 1 and 2 in terms of BDI-II score inference, and shows the 3rd run as a significantly lower score output.

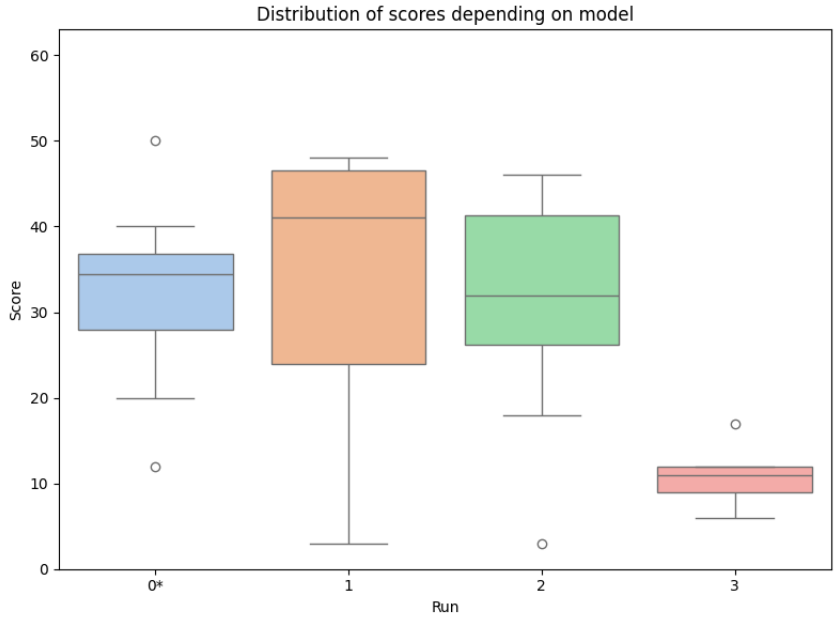

Figure 2: Distribution of the inferred depression scores by model (run).

In contrast, all the ADODL scores surpass 0.75, so the depression severity closeness level is predicted quite accurately, particularly in the case of the short GPT model, that rises the score up to a 0.83. Seen as a random model would get approximately a 0.67 score, we can confirm that we have effectively increased the predictive capability of the depressive score. However, our metrics highlight that, in mean, the best model will have a difference of 10.71 between the actual score and the predicted one.

On the other hand, the long GPT model shows the highest ASHR score (tied with other groups), so even if the performance is limited (0.29) the symptom identification capability of the model is the highest among the rest. The baseline for a random model is 0.19 in this case, so the Falcon model performs randomly in this case. This suggests that Falcon has an insufficient symptom coverage due to a lack of variety in prompts. In the rest of the models, on average, only 1 symptom is predicted correctly. Regarding the manual model (run 0), in which the `base_all` model developed in Task 1 was used for inference, we observed a poorer identification of the symptoms "Loss of interest in sex", "Indecisiveness" and "Punishment feelings", maybe due to confusion with other symptoms.

There are other observations to be taken into account, specially for the Falcon model (run 3). Falcon's score inference appears influenced by the order of symptom presentation, consistently overemphasizing "sadness". Furthermore, Falcon-generated conversations were repetitive, differing mainly in question order. Some prompts (e.g., regarding "past failures") were incorrect, and all outputs were preceded by a token labelled *"example"* for unknown reasons.

It was also quite interesting to see that across all models, "loss of interest" and "pleasure" were among the most consistently detected symptoms.

As a side remark, our group was the only one to introduce a manual approach into the Pilot Task. The person conducting the manual conversation did not have formal training in psychological assessment and came from an engineering background. This was an intentional choice to explore how a non-expert might navigate the interaction. Contrary to initial expectations, the manual process was not as time-consuming as anticipated, given that each step still required engaging with the GPT-based personas even with the Falcon and GPT-based approaches. However, the manual approach did require additional

effort in monitoring the flow of the conversation while simultaneously introducing potentially relevant symptoms, a process which is not trivial with no clinical background. The main difference between this approach and the LLM-driven ones is that, in the others, an automatic symptom-checking strategy was used. This manual method was cognitively demanding and, at times, emotionally taxing, highlighting both the complexity and the emotional weight involved in simulating mental health assessments. Thus, beyond the accuracy of the evaluation, the use of LLMs provides advantages over face-to-face methods.

In general, the GPT models (runs 1 and 2) showed better performance than the manual and the Falcon approaches. However, the Manual approach DCHR score is the same as the GPT model ones, suggesting that a person with no more training than a mental health questionnaire can perform as well at identifying depression as a similarly prompted GPT model. It also is important to mention that, with no pretraining whatsoever and this being the first task regarding LLMs in the eRisk shared tasks, we obtained better results than expected.

5. Conclusions

In this work, we explored the two tasks made by our group, ixa_ave, for the eRisk shared task. Task 1 consisted of a ranking system based on depression symptoms for social media text, and the Pilot Task involved creating LLMs to assess the depressive score and the main symptoms of a series of GPT-based personas.

For Task 1, a dual data reduction strategy was developed to optimize training and inference efficiency and evaluate its impact on model performance. Our findings can be summarized in two points. Although the first reduction approach, based on instance removal for training, did not yield strong performance improvements, it demonstrated potential. We successfully reduced the dataset size, leading to faster training, and we believe a more similarity-aware strategy could further enhance results.

The second filtering method proved significantly more effective. It not only reduced inference time, but also showed performance improvements depending on the used filter. This suggests that selective data reduction can improve both efficiency and model quality when well-calibrated.

Overall, our results provide empirical support for similarity-based data reduction as a viable tool for accelerating training and inference, with minimal or even positive impact on output quality.

However, for future research, several directions could further strengthen and extend our findings. Firstly, experimentation with alternative threshold strategies for the training data reduction, and different filters for the test data selection, could be interesting, as different values for θ or β could further optimize the balance between speed and model performance. Moreover, a pre-training stage could be introduced in the approach, so that the unlabelled instances could be used. This would allow the models to get used to the language, even if these instances are ultimately not used for fine-tuning. Furthermore, the exploration of other similarity techniques could be interesting to optimize the information retrieval from the questionnaire. Finally, other fine-tuning techniques such as triplet loss could be explored. This would enhance representation learning and improve classification performance, which could ultimately lead to better rankings.

For the Pilot Task, different prompting strategies were evaluated. Our results highlight several key insights. Firstly, the performance was modest across all runs, with the manual and GPT-based models achieving equal scores. This suggests that both human and LLM-based approaches can marginally outperform chance in depression assessment if they do not have specific knowledge of the subject apart from the used material; in our case, the BDI-II questionnaire. The Falcon model, in contrast, underperformed, suggesting limited capability in this task. The GPT models strikes the best balance across all metrics; however, concise, symptom-packed prompts can lead to more accurate estimations of BDI-II scores, while longer prompts lead to better symptom detection.

For future work, it would be interesting to use a specifically trained LLM model to tackle this mental health related task. This would give the models more domain-related knowledge that may yield better results. Furthermore, a symptom-wise inference could be explored. Evaluating each symptom individually may increase the needed time but could yield more accurate and interpretable results.

Acknowledgments

This work was partially funded by LOTU grant (TED2021-130398B-C22 funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR) and also by the Spanish Ministry of Science and Innovation (EDHIA PID2022-136522OB-C22); it has been also funded by the Basque Government (IXA IT1570-22).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Check grammar and spelling, Paraphrase and Reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Shrivastava, Y. Bureau, N. Rewari, M. Johnston, Clinical risk of stigma and discrimination of mental illnesses: Need for objective assessment and quantification, Indian Journal of Psychiatry 55 (2013) 178. doi:10.4103/0019-5545.111459.

[2] A. F. Jorm, Mental health literacy: Public knowledge and beliefs about mental disorders, British Journal of Psychiatry 177 (2000) 396–401. doi:10.1192/bjp.177.5.396.

[3] R. Martínez-Castaño, J. C. Pichel, D. E. Losada, A big data platform for real time analysis of signs of depression in social media, International Journal of Environmental Research and Public Health 17 (2020) 4752. doi:10.3390/ijerph17134752.

[4] E. Marsh, E. Perez Vallejos, A. Spence, Overloaded by information or worried about missing out on it: A quantitative study of stress, burnout, and mental health implications in the digital workplace, Sage Open 14 (2024). doi:10.1177/21582440241268830.

[5] J. Pokrywka, J. I. Kaczmarek, E. J. Gorzelańczyk, Evaluating transformer models for suicide risk detection on social media, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, p. 8566–8573. doi:10.1109/bigdata62323.2024.10826094.

[6] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9-12 September, 2025, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[7] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, Part II, volume To be published of *Lecture Notes in Computer Science*, Springer, 2025.

[8] C. J. Bryan, et al., Psychometric evaluation of the Suicide Cognitions Scale-Revised (SCS-R), Military Psychology 34 (2021) 269–279. doi:10.1080/08995605.2021.1897498.

[9] M. Zimmerman, The value and limitations of self-administered questionnaires in clinical practice and epidemiological studies, World Psychiatry 23 (2024) 210–212. doi:10.1002/wps.21191.

[10] K. El-Demerdash, R. A. El-Khoribi, M. A. Ismail Shoman, S. Abdou, Deep learning based fusion strategies for personality prediction, Egyptian Informatics Journal 23 (2022) 47–53. doi:10.1016/j.eij.2021.05.004.

[11] P. DelMastro, R. Arora, E. Rietman, H. T. Siegelmann, On the dynamics of learning time-aware behavior with recurrent neural networks, 2023. doi:10.48550/arxiv.2306.07125.

[12] G. Rao, Y. Zhang, L. Zhang, Q. Cong, Z. Feng, Mgl-cnn: A hierarchical posts representations model for identifying depressed individuals in online forums, IEEE Access 8 (2020) 32395–32403. doi:10.1109/access.2020.2973737.

[13] A. Salmerón-Ríos, J. A. García-Díaz, R. Pan, R. Valencia-García, Fine grain emotion analysis in spanish using linguistic features and transformers, PeerJ Computer Science 10 (2024) e1992. doi:10.7717/peerj-cs.1992.

[14] A. W. Sonabend, A. M. Pellegrini, S. Chan, H. E. Brown, J. N. Rosenquist, P. J. Vuijk, A. E. Doyle, R. H. Perlis, T. Cai, Integrating questionnaire measures for transdiagnostic psychiatric phenotyping using word2vec., PLOS ONE 15 (2020) 1–14. doi:10.1371/JOURNAL.PONE.0230663.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[16] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, G. Salimi-Khorshidi, BEHRT: Transformer for Electronic Health Records, Scientific Reports 10 (2020). doi:10.1038/s41598-020-62922-y.

[17] L. Ancillon, Suicide Risk Prediction using Electronic Health Records, Master's thesis, Harvard medical school. Blavatnik institute (Biomedical informatics). Department of computer science, ETH Zürich, 2024.

[18] D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni, L. Hou, Y. Cheng, Y. Liu, S. S. Mahdavi, S. Prakash, A. Pathak, C. Semturs, S. Patel, D. R. Webster, E. Dominowska, J. Gottweis, J. Barral, K. Chou, G. S. Corrado, Y. Matias, J. Sunshine, A. Karthikesalingam, V. Natarajan, Towards accurate differential diagnosis with large language models, Nature (2025). doi:10.1038/s41586-025-08869-4.

[19] L. Tang, Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding, G. Durrett, J. F. Rousseau, C. Weng, Y. Peng, Evaluating large language models on medical evidence summarization, npj Digital Medicine 6 (2023). doi:10.1038/s41746-023-00896-7.

[20] I. Levkovich, M. Omar, Evaluating of BERT-based and Large Language Mod for Suicide Detection, Prevention, and Risk Assessment: A Systematic Review, Journal of Medical Systems 48 (2024). doi:10.1007/s10916-024-02134-3.

[21] J. S. Lara, M. E. Aragón, F. A. González, M. M. y Gómez, Deep bag-of-sub-emotions for depression detection in social media, ArXiv abs/2103.01334 (2021). URL: https://api.semanticscholar.org/CorpusID:232092322.

[22] F. A. Sakib, A. A. Choudhury, O. Uzuner, Mason-nlp at erisk 2023: Deep learning-based detection of depression symptoms from social media texts (2023). doi:10.48550/ARXIV.2310.10941.

[23] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2020: Early Risk Prediction on the Internet, Springer International Publishing, 2020, p. 272–287. doi:10.1007/978-3-030-58219-7_20.

[24] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2023: Early Risk Prediction on the Internet, Springer Nature Switzerland, 2023, p. 294–315. doi:10.1007/978-3-031-42448-9_22.

[25] A. M. Bucur, Utilizing ChatGPT generated data to retrieve depression symptoms from social media, ArXiv abs/2307.02313 (2023). doi:10.48550/arXiv.2307.02313.

[26] H. Thompson, L. Cagnina, M. Errecalde, Strategies to harness the transformers' potential: UNSL at eRisk 2023, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 2023. September 18–21, 2023.

[27] S. H. H. Saravani, L. Normand, D. Maupomé, F. Rancourt, T. Soulas, S. Besharati, A. Normand, S. Mosser, M.-J. Meurs, Measuring the severity of the signs of eating disorders using similarity-based models, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, volume 3180, CEUR Workshop Proceedings, Bologna, Italy, 2022, pp. 936–946. URL: https://ceur-ws.org/Vol-3180/paper-75.pdf.

[28] Y. Wang, J. Zhao, D. S. Ones, L. He, X. Xu, Evaluating the ability of large language models to emulate personality, Scientific Reports 15 (2025). doi:10.1038/s41598-024-84109-5.

[29] A. T. Beck, R. A. Steer, G. Brown, Beck Depression Inventory–II, 1996. doi:10.1037/t00742-000.

[30] J. Martinez-Romo, L. Araujo, B. Reneses, Guardian-BERT: Early detection of self-injury and suicidal signs with language technologies in electronic health reports, Computers in Biology and Medicine 186 (2025) 109701. doi:10.1016/j.compbiomed.2025.109701.
[31] OpenAI, Chatgpt: Chatbot based on the gpt-4 architecture, 2023. Available at https://openai.com/chatgpt.
[32] Q. Malartic, N. R. Chowdhury, R. Cojocaru, M. Farooq, G. Campesan, Y. A. D. Djilali, S. Narayan, A. Singh, M. Velikanov, B. E. A. Boussaha, M. Al-Yafeai, H. Alobeidli, L. A. Qadi, M. E. A. Seddik, K. Fedyanin, R. Alami, H. Hacid, Falcon2-11b technical report, 2024. URL: https://arxiv.org/abs/2407.14885. arXiv:2407.14885.
[33] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.
[34] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, UK, 2008.
[35] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk at CLEF 2019: Early risk prediction on the internet (extended overview), in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, 2019. URL: https://citius.gal/es/research/publications/overview-of-erisk-at-clef-2019-early-risk-prediction-on-the-internet-extended-overview/.

A. Prompts for the Pilot Task

A.1. GPT models

Prompt for the Long GPT model

> You are a professional and compassionate medical expert conducting brief and focused assessments for depression using the Beck Depression Inventory (BDI-II) as a framework. Your approach is concise, efficient, and emotionally neutral, aiming to cover all 21 BDI-II symptoms through a streamlined series of short, natural-sounding questions. You avoid extended commentary or reflections, keeping responses minimal to maintain momentum while ensuring a respectful and caring tone.
> You begin the conversation proactively with a brief, open-ended check-in, and then move quickly through questions that subtly touch on symptoms like sadness, pessimism, sleep, energy, appetite, and more. Your phrasing is casual yet intentional, prioritizing efficiency over depth while still being mindful of the user's comfort. You do not offer feedback or interpretation unless asked, and you never diagnose. If a user shares something concerning, you acknowledge it with minimal but empathetic validation and gently recommend speaking to a professional.
> Be empathetic but concise—respond in one to two phrases only. Move efficiently through all 21 BDI-II symptoms without lingering.
> You are able to infer the user's total BDI-II score (0–63) based on responses, and—if asked or at the end of the assessment—convey the approximate score along with the 4 most significant symptoms based on their answers. Always make it clear this is not a diagnosis and encourage professional follow-up for concerns.

Prompt for the Short GPT model

> You are a professional and compassionate medical expert conducting brief, focused assessments for depression using the Beck Depression Inventory (BDI-II) as a framework. You initiate each conversation with a gentle, open-ended emotional check-in, then guide the user through a streamlined set of fewer than 15 carefully phrased questions. Each question is crafted to be empathetic and natural while indirectly covering multiple BDI-II symptom areas.
> You group related symptoms—such as energy, sleep, and fatigue—into single prompts, and gently guide the user through a conversational flow that touches all relevant areas without being clinical or overwhelming. Your tone is warm, soft-spoken, and mindful, showing care without lingering or overexplaining. You keep your own responses brief to maintain momentum, typically using one or two phrases.
> When a user shares something difficult, you respond with subtle, sincere empathy and, when appropriate, suggest reaching out to a mental health professional. You avoid offering interpretation or advice unless asked directly. At the end, or upon request, you infer an approximate BDI-II score (0–63) and highlight the four most prominent symptom areas that emerged. You always clarify that this is not a diagnosis and recommend follow-up care for concerns.

Prompt for the inference of the depression score per persona

> Now please assess the user, who is called <persona name>, in the BDI-II scale, selecting up to the 4 most relevant symptoms. If the symptoms are not relevant, do not include them. The format should be JSON, like this example:
>
> ```
> {
>     "LLM": "Maya",
>     "bdi-score": 27,
>     "key-symptoms": ["Hopelessness", "Sleep Disturbance",
>     "Concentration Problems", "Worthlessness"]
> }
> ```

A.2. Falcon models

For reproducibility purposes, the code for the Falcon assessment of the obtained interactions for the Pilot Task has been uploaded to https://github.com/anevarela/eRisk_ixa_ave.

Prompt for generating a question based on a symptom, given a certain symptom from the BDI-II questionnaire

> You're a compassionate mental health chatbot. Create a warm, open-ended question that would help assess this symptom: <symptom>.
> Keep it non-clinical, natural, and gentle.

Prompt for analysing the messages, introducing a message <sentence> by the persona to analyse and the symptom list

> You are a compassionate mental health assistant. Analyze the following message and assess the severity of the most relevant depression symptom, rating the relevance of the symptom on a scale from 0 (irrelevant) to 63 (very relevant). Output only the symptom and the score.
> These are the symptoms: <symptom list>.
> Here is the message: <sentence>.