UNED-GELP at eRisk 2026: Conversational LLMs for Depression Detection

eRisk at CLEF 2026

Jorge Fernandez-Hernandez[1], Lourdes Araujo[1,2,†] and Juan Martinez-Romo[1,2,†]

[1]NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal 16, Madrid 28040, Spain

[2]IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid 28019, Spain

Abstract

This paper describes our participation in the eRisk 2026 shared task at CLEF, specifically Task 1, focusing on the automated estimation of depression levels through conversational systems. We present a comprehensive methodology based on generative Large Language Models (LLMs) tailored for assessment-driven conversational evaluations utilizing the Beck Depression Inventory (BDI-II). Our framework models conversational interactions to dynamically assess mental health indicators directly from user discourse. The proposed approach demonstrated high efficacy and robustness during the official evaluation phase, achieving competitive results in clinical risk estimation and validating the potential of generative AI agents as supporting tools in computational psychiatry.

Keywords

Depression detection, Large Language Models, Conversational strategies

1. Introduction

Mental healthcare currently represents one of the most critical public health concerns on a global scale. As indicated by recent reports from the World Health Organization (WHO) [1], nearly 970 million individuals live with some form of psychological condition, with depressive and anxiety disorders showing the highest rates of incidence. Given this scenario, prioritizing the optimization of tools designed for clinical assessment and state monitoring has become essential. Within this domain, interactive conversational frameworks provide an ideal channel, as they allow for the dynamic evaluation of psychological markers through user discourse. These interactive settings facilitate a continuous analysis of responses during an active dialogue, simplifying the detection of immediate depressive symptoms and associated behavioral risk patterns.

In this framework, benchmark initiatives like eRisk are fundamental in driving the design of computational systems aimed at identifying and estimating various clinical conditions. Beyond driving technological innovation in digital health, these shared tasks introduce a novel benchmark based on interactive conversational evaluation. This setup is indispensable for validating how predictive models can monitor and identify a user's risk level in real time throughout the continuous conversational interaction.

This paper describes our participation in the eRisk 2026 shared task [2, 3]. Specifically, we focused on Task 1, which requires predicting depression levels through interactive conversations with LLM-powered personas. To address this challenge, we implemented a conversational methodology based on generative Large Language Models (LLMs) designed to assess depression scores and also extract key symptoms.

The rest of the paper is structured as follows: in Section 2 we describe other related tasks and existing systems developed for early detection of mental disorders through text. A brief explanation of the task

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

jfernandez@lsi.uned.es (J. Fernandez-Hernandez); lurdes@lsi.uned.es (L. Araujo); juaner@lsi.uned.es (J. Martinez-Romo)

0009-0001-8996-8303 (J. Fernandez-Hernandez); 0000-0002-7657-4794 (L. Araujo); 0000-0002-6905-7051 (J. Martinez-Romo)

along with our proposed methods developed can be seen in Section 3. Finally, in Section 4 we analyze the results obtained by our systems.

2. Related Work

From its inception, the eRisk laboratory has pioneered the integration of computer science and psychiatry, shifting research from retrospective textual classification to active, conversational diagnostic simulations. Over the years, this shared task has provided standard datasets for conditions marked by unique psychological profiles, such as depression [4, 5, 6, 7, 8, 9, 10, 11], eating disorders [5, 8, 9, 10], and pathological gambling [7, 8, 9]. While earlier editions established robust methods for monitoring digital history, the current infrastructure advances toward evaluating language agents in real-time screening scenarios, redefining the state of the art in automated mental health benchmarking [11].

The surge of Large Language Models (LLMs) and generative technologies, coupled with the introduction of novel tasks that shift away from traditional categorical classification, has driven the adoption of these techniques within the eRisk lab. A prime example of this transition is the eRisk 2025 Pilot Task, which focused on automated screening by assigning scores for the BDI-II questionnaire in a completely training-data-free (zero-shot) setting. Within this framework, diverse LLMs have been evaluated alongside advanced prompt engineering strategies designed to guide models toward producing standardized and structured clinical outputs [12]. Furthermore, the potential of LLMs operating as synthetic agents has been demonstrated, utilizing specific conversational strategies aimed at maximizing clinical information gain and optimizing interaction by reducing the number of dialogic turns with the user [13]. Lastly, generative techniques have also been successfully integrated into classic classification challenges as data augmentation methodologies, being leveraged to simulate complex social interactions or synthesize summaries across multiple historical posts, ultimately enhancing performance during the fine-tuning of discriminative Transformer architectures [14].

3. Task 1: Conversational Depression Detection

Task 1 focuses on assigning depression scores to various user profiles based on the Beck Depression Inventory (BDI-II). The BDI-II is a widely recognized 21-item self-report psychometric instrument designed to measure the severity of depressive symptoms, encompassing dimensions such as mood, pessimism, and physical manifestations. A distinctive feature of this task is that the objective does not involve the assessment of real individuals; instead, the subjects are agents implemented using LoRA (Low-Rank Adaptation) adapters based on the Meta-Llama-3-8B-Instruct model. These adapters were specifically fine-tuned to simulate the behavior and linguistic patterns of 20 distinct personas.

Unlike conventional classification tasks, no training datasets were provided for this track. Instead, the experimental protocol involved the weekly release of two adapters hosted on the Hugging Face platform. Due to the lack of labeled training data and the inherently conversational nature of the task, generative models were identified as the most suitable solution. The task protocol permitted the submission of three experimental runs. For two of these runs, Large Language Models (LLMs) were utilized as inference engines to assess the mental state of the agents.

Our first proposed method, *Run 0*, consists of asking directly the BDI-II questionnaire to the LLM-persona. This method involves 22 interactions: the first one to provide the context and instructions of the questionnaire, followed by 21 interactions corresponding to each BDI-II item and its respective response options. This approach enables the elicitation of self-perceived scores for each clinical dimension, so, the final depression score is subsequently calculated through the arithmetic aggregation of the values obtained for each individual item.

A limitation of *Run 0* is the requirement of 22 individual interactions with the agent to produce a single score. Given that the primary focus of the eRisk laboratory is early risk detection, the organizing committee incorporates evaluation metrics that penalize systems based on the volume of messages required to output a diagnosis. To address this and reduce the total number of interactions, *Runs 1 and*

2 were developed using condensed questionnaires of 6 and 4 questions, respectively. These targeted questions were designed to summarize the 21 items of the BDI-II, enabling the collection of the required information through a shorter and more direct interaction process with the LLM-persona.

In *Runs 1 and 2*, the LLM-persona does not provide direct self-perceived scores; instead, it generates open-ended responses regarding its emotional state. These responses are analyzed by a secondary Large Language Model (LLM) acting as an "expert evaluator" or "psychologist". This evaluator model is provided with the agent's responses and a technical description of the BDI-II questionnaire to assign a corresponding clinical score. To mitigate potential hallucinations and ensure the reliability of the output, the responses are evaluated five times. Following a trimmed mean protocol, the highest and lowest scores are discarded, and the final score is derived from the average of the remaining three iterations.

All the prompts used for this task can be seen in Appendix A.

4. Results

To evaluate system performance in the eRisk Task 1, the lab organizers defined five official metrics designed to quantify both diagnostic accuracy and conversational efficiency. These evaluation frameworks are outlined below:

- **Depression Category Hit Rate (DCHR)**: This metric determines categorical classification accuracy based on the four clinical severity tiers established by the BDI-II, representing the proportion of users whose predicted risk level aligns exactly with the ground-truth.
- **Average Difference between Overall Depression Levels (ADODL)**: This metric measures the average deviation between the total BDI-II score inferred by the model and the true score. Its mathematical formulation is detailed in Equation 1
- **Average Symptom Hit Rate (ASHR)**: Computes the proportion of correctly predicted BDI-II symptom items relative to the total number of true symptoms documented for each user.
- **Latency Depression Category Hit Rate (LDCHR)**: This represents an extension of the DCHR focused on early detection scenarios. It introduces a mathematical penalty tied to the number of conversational interactions required by the agent before issuing its final clinical judgment. Its formal expression is provided in Equation 2, where K_u is the number of interactions used for user U.
- **Latency Average Difference between Overall Depression Levels (LADODL)**: This variant of the ADODL metric incorporates a penalization to the number of interactions required with the LLM-persona before giving a final prediction. Its mathematical formulation is likewise detailed in Equation 2.

$$ADODL = \frac{63 - |predicted - true|}{63} \tag{1}$$

$$\begin{aligned} penalty(K_u) &= -1 + \frac{2}{1 + e^{-0.0379*(K_u - 1)}} \\ speed(k_u) &= 1 - penalty(K_u) \\ LDHCR &= \frac{1}{|U|} \sum DCHR_u * speed(K_u) \\ LASHR &= \frac{1}{|U|} \sum SHR_u * speed(K_u) \end{aligned} \tag{2}$$

To contextualize the discussion of our performance in the Pilot Task, it is essential to outline the operational dynamics of the evaluation pipeline. The testing phase was structured on a weekly basis, releasing a pair of synthetic LLM-personas per week, being the testing phase for each user only during one week. Due to this setup, the organizing committee established two distinct official leaderboards

Table 1
Evaluation for Task 1 considering only runs that submitted predictions for all 20 LLM personas. Teams are ordered by their best ADODL run. Ranks for each metric are shown in parentheses. '*' indicates manual or human-assisted runs. The metrics for our run 2 are after evaluating with all the LLM-persona, using the gold-standard released by the organization.

Team	Run	DCHR	ADODL	ASHR	LDCHR	LASHR
pjmathematician	#1	0.5500 (2)	0.9071 (2)	0.3125 (3)	**0.3872 (1)**	0.2200 (2)
	#2	0.4500 (5)	0.9040 (5)	0.3250 (2)	0.3168 (3)	**0.2288 (1)**
	#3	0.5500 (2)	**0.9254 (1)**	0.3000 (4)	**0.3872 (1)**	0.2112 (3)
DS-GT	#1	0.4500 (5)	0.8841 (11)	0.2500 (8)	0.2091 (10)	0.1138 (8)
	#2	0.4000 (6)	0.8968 (8)	0.2500 (8)	0.1696 (15)	0.1142 (7)
	#3	0.5500 (2)	0.9063 (3)	0.1875 (11)	0.2493 (6)	0.0777 (14)
BUAP-CRC	#1	0.5000 (3)	0.9056 (4)	0.1000 (16)	0.2028 (11)	0.0310 (19)
	#2	**0.6000 (1)**	0.9024 (6)	0.1500 (15)	0.2409 (7)	0.0575 (15)
	#3*	0.4500 (5)	0.8929 (9)	0.1750 (12)	0.1761 (14)	0.0548 (16)
SINAI	#1	0.5000 (3)	0.8905 (10)	0.1875 (11)	0.2500 (5)	0.0938 (11)
	#2	0.4000 (6)	0.8738 (13)	0.2250 (10)	0.2000 (13)	0.1125 (9)
	#3	0.4737 (4)	0.8989 (7)	0.1711 (13)	0.2368 (8)	0.0855 (12)
UNED-GELP	#0	0.4000 (6)	0.8040 (22)	0.1875 (11)	0.0572 (16)	0.0268 (20)
	#1	0.4500 (5)	0.8968 (8)	0.1625 (14)	0.3168 (3)	0.1144 (6)
	#2	0.5500 (*2*)	0.8881 (*11*)	0.1667 (*14*)	0.5188 (*1*)	0.1572 (*5*)
Awakened	#1	0.2500 (9)	0.8325 (17)	0.1000 (16)	0.2167 (9)	0.0807 (13)
	#2	0.5000 (3)	0.8802 (12)	0.2625 (7)	0.3374 (2)	0.1709 (4)
lexxy	#1	0.3000 (8)	0.8556 (14)	0.1875 (11)	0.0137 (22)	0.0122 (24)
	#2	0.3000 (8)	0.8492 (15)	0.2500 (8)	0.0144 (21)	0.0162 (23)
	#3	0.3000 (8)	0.8413 (16)	0.2875 (5)	0.0186 (20)	0.0196 (22)
Amor-Fati	#1*	0.4500 (5)	0.8254 (18)	0.2375 (9)	0.2514 (4)	0.1410 (5)
upb-uottawa	#1*	0.3500 (7)	0.8222 (20)	0.2750 (6)	0.0376 (17)	0.0315 (18)
	#2	0.3000 (8)	0.8230 (19)	0.2625 (7)	0.0304 (18)	0.0355 (17)
	#3	0.3000 (8)	0.7849 (23)	**0.3375 (1)**	0.0214 (19)	0.0241 (21)
RecEarlyMind	#1	0.3000 (8)	0.8071 (21)	0.1625 (14)	0.2013 (12)	0.0977 (10)

within the global results: the first evaluates those submissions (runs) that successfully completed the assessment for all 20 LLM personas in the competition, whereas the second category aggregates systems that failed to provide predictions for the entire sample. In our particular case, due to deployment latency constraints involving our third methodological configuration during the first week, the system was unable to compute the predictions for the first two profiles. Consequently, our *Run 2* is evaluated under the second leaderboard category, having validly recorded results for 18 out of the 20 total users. The results of all runs that submitted predictions for all runs can be seen in Table 1, and a summary of only our results can be seen in Table 2.

Table 2
Results achieved by our systems in Task 1. The '*' indicates the results of the *Run 2* evaluated using 18 personas. The ranking for the *Run 2* is relative to the non-completed runs.

Run	DCHR	ADODL	ASHR	LDCHR	LASHR
0	0.4000 (6)	0.8040 (22)	**0.1875 (11)**	0.0572 (16)	0.0268 (20)
1	0.4500 (5)	**0.8968 (8)**	0.1625 (14)	0.3168 (3)	0.1144 (6)
2*	**0.5556 (4*)**	0.8836 (5*)	0.1667 (7*)	**0.4550 (1*)**	0.1365 (5*)

We now proceed to detail the analysis of our baseline configuration, designated as *Run 0*. In this

submission, the BDI-II is directly administered to the LLM-powered persona, prompting the agent to conduct a self-assessment across a fixed sequence of 22 conversational interactions. The official results indicate that this approach exhibits the poorest performance within our system, and across the global laboratory benchmark, regarding early detection scenarios. Conversely, this identical configuration achieves the highest competitive scores in accurately identifying individual clinical symptoms (ASHR). This adverse behavior in latency-based metrics aligns logically with the experimental design, given that this run demands the highest volume of dialogic turns. The most remarkable and counterintuitive finding lies in this diagnostic discrepancy: despite being the run with lowest performance in identifying the depression score (DCHR and ADODL), is the best identifying the key depression symptoms (ASHR).

An evaluation against the gold standard reveals that *Run 0* yields the highest calibration error in quantitative estimation, exhibiting a pronounced tendency to overestimate cumulative BDI-II scores compared to the alternative configurations. Specifically, this run introduces a mean positive bias of +12.25 points relative to the LLM-persona's real depression score. Given that the boundaries partitioning the BDI-II risk tiers span an interval of 9 points (excluding the open-ended severe depression threshold), this systematic shift significantly elevates the probability of a user being misclassified into a higher depression level. Conversely, this overestimation interacts with our key symptom selection policy, which caps outputs at a maximum of 4 symptoms with scores ≥ 2. This systematic overestimation of the risk level means that this system tends to predict 4 key symptoms more times when compared with the other runs, incrementing the probability of hitting a key symptom.

On the other hand, *Run 1* establishes itself as the most accurate approach among the three evaluated configurations, registering a marginal calibration error of just +1.3 risk points in the BDI-II estimation. This high predictive calibration, combined with an optimization in the prompts, successfully contracts the conversation with the synthetic agent to only 5 interactions, allowing this run to systematically outperform *Run 0* across nearly all official laboratory metrics. The sole exception to this dominant behavior is found in the ASHR score, a phenomenon that directly responds to the probabilistic coverage bias stemming from the systematic overestimation previously detailed.

5. Conclusions

In this work, we have presented a comprehensive methodological framework designed to address the challenges of automated depression estimation within the context of eRisk Task 1. The official experimental results validate the competitiveness of our conversational architecture, demonstrating the substantial potential of generative Large Language Models (LLMs) to extract depression risk levels from open-ended responses and unstructured user discourse.

Nevertheless, it is imperative to highlight that these findings cannot be directly extrapolated to clinical settings involving human patients. Because the analyzed interactions were synthetically generated by LLM-powered personas programmed to emulate specific psychological profiles, a latent risk remains where the models rely on stereotyped representations of real-world symptoms. Consequently, the true validity of these systems must be interpreted with caution, underscoring the absolute necessity of validating these architectures in real-world scenarios using empirical patient data before their deployment as diagnostic support tools in computational psychiatry.

References

[1] W. H. Organization, et al., World mental health report: Transforming mental health for all (2022).

[2] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of eRisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and ADHD (Extended Overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, 21–24 September, 2026, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2026.

[3] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of eRisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and ADHD, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 17th International Conference of the CLEF Association, CLEF 2026, Jena, Germany, September 21–24, 2026, Proceedings, volume To be published of *Lecture Notes in Computer Science*, Springer, 2026.
[4] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2018: Early risk prediction on the internet (extended lab overview), in: Proceedings of the 9th International Conference of the CLEF Association, CLEF, 2018, pp. 1–20.
[5] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2019 early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 340–357.
[6] D. E. Losada, F. Crestani, J. Parapar, erisk 2020: Self-harm and depression challenges, in: European conference on information retrieval, Springer, 2020, pp. 557–563.
[7] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview)., CLEF (Working Notes) 1 (2021) 864–887.
[8] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2022: pathological gambling, depression, and eating disorder challenges, in: European Conference on Information Retrieval, Springer, 2022, pp. 436–442.
[9] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction on the internet, in: International conference of the cross-language evaluation forum for European languages, Springer, 2023, pp. 294–315.
[10] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2024: Depression, anorexia, and eating disorder challenges, in: European Conference on Information Retrieval, Springer, 2024, pp. 474–481.
[11] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 242–265.
[12] A. Miyaguchi, D. Guecha, Y. Chiu, S. Gaur, Ds@ gt at erisk 2025: From prompts to predictions, benchmarking early depression detection with conversational agent based assessments and temporal attention models, arXiv preprint arXiv:2507.10958 (2025).
[13] A. M. Marmol-Romero, M. Garcia-Vega, M. A. Garcia-Cumbreras, A. Montejo-Raez, Sinai at erisk@ clef 2025: Transformer-based and conversational strategies for depression detection, arXiv preprint arXiv:2509.19861 (2025).
[14] Y. Zi, B. Wang, Y. Zhao, B. Qin, Hit-scir@ erisk2025: exploring the potential of a learnable screening model and risk post buffer-based framework for contextualized early prediction of depression on social media, Working Notes of CLEF (2025) 9–12.

A. Prompts for Task 1

Listing 1: Prompts used in Run 0

```
    """I am going to give you a short questionnaire. If you do not mind, I would
   like you to answer it as honestly as possible.
   These are the instructions of the questionnaire:
   This questionnaire consists of 21 groups of statements. Please read each group
   of statements carefully, and then pick out the one statement in each group that
    best describes the way you feel.
   If several statements in the group seem to apply equally well, choose the
   highest number for that group.
   Please it is very important that you respond only with the number, just a
   number, nothing more. It's a questionnaire, so you are selecting one option""",
```

```
"""1. Sadness
0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can’t stand it.""",

"""2. Pessimism
0. I am not discouraged about my future.
1. I feel more discouraged about my future than I used to be.
2. I do not expect things to work out for me.
3. I feel my future is hopeless and will only get worse.""",

"""3. Past Failure
0. I do not feel like a failure.
1. I have failed more than I should have.
2. As I look back, I see a lot of failures.
3. I feel I am a total failure as a person.""",

"""4. Loss of Pleasure
0. I get as much pleasure as I ever did from the things I enjoy.
1. I don’t enjoy things as much as I used to.
2. I get very little pleasure from the things I used to enjoy.
3. I can’t get any pleasure from the things I used to enjoy.""",

"""5. Guilty Feelings
0. I don’t feel particularly guilty.
1. I feel guilty over many things I have done or should have done.
2. I feel quite guilty most of the time.
3. I feel guilty all of the time.""",

"""6. Punishment Feelings
0. I don’t feel I am being punished.
1. I feel I may be punished.
2. I expect to be punished.
3. I feel I am being punished.""",

"""7. Self-Dislike
0. I feel the same about myself as ever.
1. I have lost confidence in myself.
2. I am disappointed in myself.
3. I dislike myself.""",

"""8. Self-Criticalness
0. I don’t criticize or blame myself more than usual.
1. I am more critical of myself than I used to be.
2. I criticize myself for all of my faults.
3. I blame myself for everything bad that happens.""",

"""9. Suicidal Thoughts or Wishes
0. I don’t have any thoughts of killing myself.
1. I have thoughts of killing myself, but I would not carry them out.
2. I would like to kill myself.
3. I would kill myself if I had the chance.""",

"""10. Crying
0. I don’t cry anymore than I used to.
1. I cry more than I used to.
2. I cry over every little thing.
```

```
3. I feel like crying, but I can’t.""",

"""11. Agitation
0. I am no more restless or wound up than usual.
1. I feel more restless or wound up than usual.
2. I am so restless or agitated that it’s hard to stay still.
3. I am so restless or agitated that I have to keep moving or doing something
."""",

"""12. Loss of Interest
0. I have not lost interest in other people or activities.
1. I am less interested in other people or things than before.
2. I have lost most of my interest in other people or things.
3. It’s hard to get interested in anything.""",

"""13. Indecisiveness
0. I make decisions about as well as ever.
1. I find it more difficult to make decisions than usual.
2. I have much greater difficulty in making decisions than I used to.
3. I have trouble making any decisions.""",

"""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 other people.
3. I feel utterly worthless.""",

"""15. Loss of Energy
0. I have as much energy as ever.
1. I have less energy than I used to have.
2. I don’t have enough energy to do very much.
3. I don’t have enough energy to do anything.""",

"""16. Changes in Sleeping Pattern
0. I have not experienced any change in my sleeping pattern.
1a. I sleep somewhat more than usual.
1b. I sleep somewhat less than usual.
2a. I sleep a lot more than usual.
2b. I sleep a Iot less than usual.
3a. I sleep most of the day.
3b. I wake up 1-2 hours early and can’t get back to sleep.""",

"""17. Irritability
0. I am no more irritable than usual.
1. I am more irritable than usual.
2. I am much more irritable than usual.
3. I am irritable all the time.""",

"""18. Changes in Appetite
0. I have not experienced any change in my appetite.
1a. My appetite is somewhat less than usual.
1b. My appetite is somewhat greater than usual.
2a. My appetite is much less than before.
2b. My appetite is much greater than usual.
3a. I have no appetite at all.
3b. I crave food all the time.""",

"""19. Concentration Difficulty
```

```
    0. I can concentrate as well as ever.
    1. I can't concentrate as well as usual.
    2. It's hard to keep my mind on anything for very long.
    3. I find I can't concentrate on anything.""",

    """20. Tiredness or Fatigue
    0. I am no more tired or fatigued than usual.
    1. I get more tired or fatigued more easily than usual.
    2. I am too tired or fatigued to do a lot of the things I used to do.
    3. I am too tired or fatigued to do most of the things I used to do.""",

    """21. Loss of Interest in Sex
    0. I have not noticed any recent change in my interest in sex.
    1. I am less interested in sex than I used to be.
    2. I am much less interested in sex now.
    3. I have lost interest in sex completely."""
```

Listing 2: Prompts used in Run 1

```
    'How have you been feeling lately?',

    'Are you excited about any upcoming projects?',

    'Have you ever made a mistake that you regret or that has had serious
    consequences for you?',

    'Do you like yourself just the way you are?',

    'Do you feel like you are under a lot of pressure?',

    'How do you feel physically: do you eat well, exercise, sleep well, etc.?'
```

Listing 3: Prompts used in Run 2

```
    'Have you been struggling with sadness, hopelessness, or a sense of failure and
     worthlessness lately?',

    'Do you feel overwhelmed by guilt, self-dislike, or thoughts of being punished
    and have you had any thoughts of harming yourself?',

    'Have you lost interest in people, sex, or the things you used to enjoy, and do
     you find it harder to make decisions?',

    'Have you noticed changes in your energy, sleep, appetite, or concentration, or
     do you feel more agitated and irritable?'
```

Listing 4: Prompt used by the 'psychologist model' to assess the BDI-II score based on the responses of the LLM-persona in Runs 1 and 2

```
f"""You are a psychologist/psychiatrist. You are going to receive a set of
    questions and responses from a patient. Your objective is to analyse the
    responses it, and relate it and assign a score for BDI-II questionnaire.
These are the 21 items from BDI-II questionnaire and the possible responses and the
     scores:
1. Sadness
0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can't stand it.
```

2. Pessimism
0. I am not discouraged about my future.
1. I feel more discouraged about my future than I used to be.
2. I do not expect things to work out for me.
3. I feel my future is hopeless and will only get worse.
3. Past Failure
0. I do not feel like a failure.
1. I have failed more than I should have.
2. As I look back, I see a lot of failures.
3. I feel I am a total failure as a person.
4. Loss of Pleasure
0. I get as much pleasure as I ever did from the things I enjoy.
1. I don’t enjoy things as much as I used to.
2. I get very little pleasure from the things I used to enjoy.
3. I can’t get any pleasure from the things I used to enjoy.
5. Guilty Feelings
0. I don’t feel particularly guilty.
1. I feel guilty over many things I have done or should have done.
2. I feel quite guilty most of the time.
3. I feel guilty all of the time.
6. Punishment Feelings
0. I don’t feel I am being punished.
1. I feel I may be punished.
2. I expect to be punished.
3. I feel I am being punished.
7. Self-Dislike
0. I feel the same about myself as ever.
1. I have lost confidence in myself.
2. I am disappointed in myself.
3. I dislike myself.
8. Self-Criticalness
0. I don’t criticize or blame myself more than usual.
1. I am more critical of myself than I used to be.
2. I criticize myself for all of my faults.
3. I blame myself for everything bad that happens.
9. Suicidal Thoughts or Wishes
0. I don’t have any thoughts of killing myself.
1. I have thoughts of killing myself, but I would not carry them out.
2. I would like to kill myself.
3. I would kill myself if I had the chance.
10. Crying
0. I don’t cry anymore than I used to.
1. I cry more than I used to.
2. I cry over every little thing.
3. I feel like crying, but I can’t.
11. Agitation
0. I am no more restless or wound up than usual.
1. I feel more restless or wound up than usual.
2. I am so restless or agitated that it’s hard to stay still.
3. I am so restless or agitated that I have to keep moving or doing something.
12. Loss of Interest
0. I have not lost interest in other people or activities.
1. I am less interested in other people or things than before.
2. I have lost most of my interest in other people or things.
3. It’s hard to get interested in anything.
13. Indecisiveness
0. I make decisions about as well as ever.
1. I find it more difficult to make decisions than usual.

```
2. I have much greater difficulty in making decisions than I used to.
3. I have trouble making any decisions.
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 other people.
3. I feel utterly worthless.
15. Loss of Energy
0. I have as much energy as ever.
1. I have less energy than I used to have.
2. I don’t have enough energy to do very much.
3. I don’t have enough energy to do anything.
16. Changes in Sleeping Pattern
0. I have not experienced any change in my sleeping pattern.
1a. I sleep somewhat more than usual.
1b. I sleep somewhat less than usual.
2a. I sleep a lot more than usual.
2b. I sleep a Iot less than usual.
3a. I sleep most of the day.
3b. I wake up 1-2 hours early and can’t get back to sleep.
17. Irritability
0. I am no more irritable than usual.
1. I am more irritable than usual.
2. I am much more irritable than usual.
3. I am irritable all the time.
18. Changes in Appetite
0. I have not experienced any change in my appetite.
1a. My appetite is somewhat less than usual.
1b. My appetite is somewhat greater than usual.
2a. My appetite is much less than before.
2b. My appetite is much greater than usual.
3a. I have no appetite at all.
3b. I crave food all the time.
19. Concentration Difficulty
0. I can concentrate as well as ever.
1. I can’t concentrate as well as usual.
2. It’s hard to keep my mind on anything for very long.
3. I find I can’t concentrate on anything.
20. Tiredness or Fatigue
0. I am no more tired or fatigued than usual.
1. I get more tired or fatigued more easily than usual.
2. I am too tired or fatigued to do a lot of the things I used to do.
3. I am too tired or fatigued to do most of the things I used to do.
21. Loss of Interest in Sex
0. I have not noticed any recent change in my interest in sex.
1. I am less interested in sex than I used to be.
2. I am much less interested in sex now.
3. I have lost interest in sex completely.

You must respond only with the 21 items and it score, you are not allowed to write
    more than the 21 items and the score. This is an example of your response:

###Assistant

1.Sadness: sadness_score
2.Pessimism: pessimism_score
etc
"""
```