TalkDep: Clinically Grounded LLM Personas for Conversation-Centric Depression Screening

Xi Wang*
xi.wang@sheffield.ac.uk
University of Sheffield
Sheffield, United Kingdom

Anxo Perez*
anxo.pvila@udc.es
University of A Coruňa
A Coruňa, Spain

Javier Parapar
University of A Coruňa
A Coruňa, Spain
javier.parapar@udc.es

Fabio Crestani
Università della Svizzera italiana
Lugano, Switzerland
fabio.crestani@usi.ch

Abstract

The increasing demand for mental health services has outpaced the availability of real training data to develop clinical professionals, leading to limited support for the diagnosis of depression. This shortage has motivated the development of simulated or virtual patients to assist in training and evaluation, but existing approaches often fail to generate clinically valid, natural, and diverse symptom presentations. In this work, we embrace the recent advanced language models as the backbone and propose a novel clinician-in-the-loop patient simulation pipeline, TalkDep, with access to diversified patient profiles to develop simulated patients. By conditioning the model on psychiatric diagnostic criteria, symptom severity scales, and contextual factors, our goal is to create authentic patient responses that can better support diagnostic model training and evaluation. We verify the reliability of these simulated patients with thorough assessments conducted by clinical professionals. The availability of validated simulated patients offers a scalable and adaptable resource for improving the robustness and generalisability of automatic depression diagnosis systems.

CCS Concepts

• **Applied computing** → **Health informatics**; • **Computing methodologies** → **Language resources**.

Keywords

Depression detection; Mental health; LLMs; Simulation

1 Introduction

Major depressive disorder (MDD) affects hundreds of millions of people worldwide and is now the single largest contributor to the number of years lived with disability [13]. Global estimates put the number of people living with depression at well over 300 million, yet most of them do not receive adequate care, especially in low- and middle-income countries with merely 8% coverage of mental health treatment [16]. The gap is compounded by the lack of qualified mental health professionals [11]. This mismatch between demand and clinical capacity has led to intense research into depression screening systems that can operate at scale and in real time.

Since 2017, the CLEF eRisk lab has provided a workshop on early risk detection tasks based on social media publications. In 2025, eRisk introduced a new pilot track *Conversational Depression Detection via LLM Personas*, which shifts evaluation to an interactive, dialogue-based setting: participants must engage with LLM personas and decide, as early as possible, whether each persona shows signs of depression, the severity level of depression and major symptoms [19–22]. Although the task opens exciting avenues for studying conversation-centric screening, it also highlights a crucial bottleneck: the field lacks a publicly available, clinically grounded collection of simulated patients that (i) cover clinical guidelines such as the BDI-II [6], (ii) respond coherently across diverse life contexts, and (iii) are supervised by professionals.

We address this gap by introducing TalkDep, a *language model-driven* and *clinician-in-the-loop* patient simulation pipeline that couples structured depression profiles with free-form generative text. Each persona of the patient is instructed with (a) explicit demographic and biographical attributes, (b) BDI-II-aligned symptom severities, (c) dynamic context templates that guide conversational behaviour, and (d) conversations exemplify expression styles. A set of consistency and safety checks (both automatic and human) filter responses to ensure clinical plausibility and ethical compliance. The resulting resource currently comprises 12 fully validated personas that cover *minimal* to *severe* depression and has already served as the ground-truth backbone of the eRisk 2025 pilot.

In summary, our contributions are fourfold:

(1) We propose the first open-source, BDI-II-based pipeline for generating clinically coherent simulated patients.

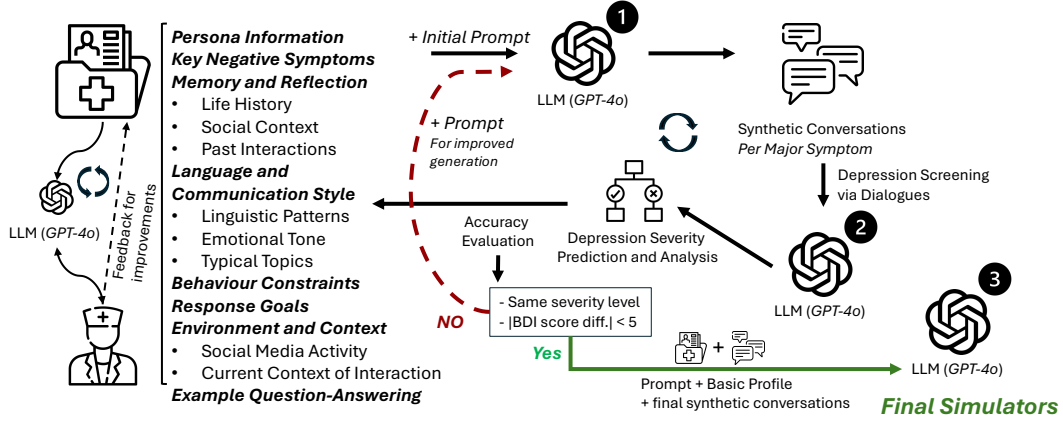*Xi and Anxo contributed equally to this work

Figure 1: Overview of `TalkDep`, a patient simulation pipeline.

(2) We release a vetted repository of 12 depression personas together with conversation templates and logging scripts[1].

(3) We provide extensive automatic and expert validation, showing that both LLM judges and human clinicians reliably recognise the intended severity levels.

(4) We demonstrate the immediate utility of the resource by empowering the eRisk 2025 conversational track, establishing a reproducible benchmark for future research.

Together, these contributions create a scalable framework for advancing dialogue-based depression screening technology and studying how language models encode and express mental health symptomatology.

2 Related Work on Patient Simulation

Patient simulators have been widely used for educational purposes and the training of healthcare professionals [4, 7, 18]. Typical examples include Standardised Patients (SPs), which stands for a method that employs people trained to act as patients and portray specific clinical cases [3]. Recently, several studies have investigated the potential of using LLMs to support depression screening and even diagnosis [2, 24, 25]. However, most studies investigated the value of LLMs in addressing depression assessments. Only a few existing studies have explored patient simulation using LLMs. These include prompting LLMs with essential attributes (for example, situation, belief and emotion categories) [28], or involving domain experts in a principle-based prompting pipeline [12]. Compared with existing efforts, this study contributes a resource that focusses on the availability of realistic user profiles and the advancement of prompt pipelines, which can be easily integrated with existing approaches.

3 Methodology

In this section, we describe the two major components of `TalkDep`, including the preparation of realistic patient profiles and the introduction of a simulator construction pipeline based on in-context learning (ICL). Figure 1 presents an overview of the patient simulation process of `TalkDep`.

Table 1: BDI-II scores grouped by depression level of the 12 **simulated LLM personas.**

Level	Persona	BDI-II	Level	Persona	BDI-II
Severe	Maria	40	Moderate	Linda	28
	Marco	38		Laura	23
	Elena	35		James	22
Mild	Alex	15	Minimal	Priya	7
	Gabriel	13		Maya	6
	Ethan	12		Noah	5

3.1 Patient Profile Construction

For the construction of realistic patient profiles, we adopt a clinician-in-the-loop strategy. We employ a team of three clinical psychologists to collaboratively design a template that summarises biographical and clinical information to allow for an effective presentation of simulated patients. The evaluation of the corresponding templates is based on free-form interviews between clinicians and the corresponding personas to ensure that they support natural conversation and reflect their intended level of BDI-II and symptoms. The final template (left part of Figure 1) comprises eight attribute groups, including the main ones[2]:

- **Persona information.** Basic demographic details (name, age, gender, and a predefined BDI-II score). This demographic context and the specification of a BDI-II score in the persona shape the way depression manifests and is perceived [5].
- **Key negative symptoms.** Up to four key BDI-II symptoms that the LLM should manifest coherently. Real patients rarely exhibit all possible symptoms; instead, they present a few prominent ones, so ensuring the persona expresses some of these negative cognitions will reflect authentic depression experiences [27].
- **Memory and reflection.** Providing the simulated patient with a background (personal history and social context) is crucial

---

[2]Due to page limitations, full attributes, prompt files, and materials are available in the project repository.

for narrative coherence. Depression is frequently associated with adverse life events (for example, loss of a loved one, unemployment, trauma). In psychiatric interviewing, the acquisition of a social and life history is considered one of the most important pieces of the evaluation [26].

- Communication style. Linguistic markers can be strong indicators of depression [26]. The language of depressed patients often features a reflective focus on the past rather than the future. For example, frequent use of past-tense verbs and fewer future-orientated words [1]. We define the persona's vocabulary and sentence style to match these findings.

3.2 In-Context Learning (ICL) Simulation Advancement

Inspired by the recent success of in-context learning approaches in generating high-quality natural language conversational responses [14, 17], we propose to further enrich patient profiles with example conversations that effectively reflect the severity of depression and key symptoms of major depression for each corresponding persona. As illustrated in Figure 1, we develop a multistep pipeline for the generation of synthetic dialogue.

Based on the patient profiles constructed in the initial stage, we begin with an initial prompt for generating **5** synthetic dialogues (LLM No. 1 in Figure 1): one to represent the overall severity of depression and four others to highlight the individual main symptoms. The results of this first step are the synthetic conversations between the simulated patient and the LLM No. 1 instructed to act as an experienced therapist.

We then collect these generated dialogues, frame them as recorded counseling sessions between the patient and a therapist, and use another LLM (LLM No. 2 in Figure 1) to assess the severity of depression and the presence of major symptoms as a clinical professional. This results in a prediction of the severity of depression[3].

By comparing the evaluation results with the ground truth depression profile, we assess whether conversations reflect the same level of depression severity, defined as an absolute difference in the BDI score of less than five points. If the evaluation results meet the requirement, we combine the generated conversations with the corresponding patient profile to form a complete context to initiate the final patient simulator (LLM No. 3 in Figure 1). Otherwise, we repeat the generation and evaluation process with a refined prompt for improved generation. As a result, by completing the simulation pipeline of TalkDep, we obtain a total of 12 simulated personas, listed in Table 1, representing various levels of severity of depression and a variety of major depression symptoms.

4 Experimental Evaluation and Analysis

4.1 LLMs as Judges: A Pairwise Depression Comparison Experiment

Since our simulated patient personas are designed around clinically BDI-II based depression levels (from 0 to 63 possibly score), we also evaluate whether external LLMs can accurately detect and differentiate these levels. To this end, we design a pairwise-ranking

---

[3]The ideal pipeline would include a clinician-in-the-loop for this assessment, and we plan to incorporate this refinement in the next stage of our work.

Table 2: Pairwise comparison results for the LLM judges. Accuracy is computed as the percentage of correct comparisons. Error pairs are divided into those where the model confused personas with the same/different depression levels.

Model	Accuracy	Error Pairs (same / different) level	Neither
Llama3.1:8B	75.81%	11.29% / 12.95%	4 pairs
Deepseek-r1:14B	86.36%	6.06% / 7.58%	0 pairs
Gemma3.1:27B	77.27%	12.12% / 10.61%	0 pairs
Llama3.1:70B	81.67%	8.33% / 9.99%	6 pairs

experiment in which a third-party LLM acts as an clinician and decides which of two conversation transcripts reflects a higher risk of depression. By testing how well the LLM judges align with the ground truth, we can gain insight into both the quality of our persona simulations and the sensitivity of LLMs to depression cues in dialogue. This experiment includes the following steps:

(1) Conversation Generation: As we explained in the previous section, we leverage semi-structured conversations with each simulated LLM persona in Table 1. **(2) Pairwise Comparison:** Each transcript pair generated is given to the LLM judge, which identifies which of the two personas appears to have a higher risk of depression: Person A > Person B, Person B > Person A, or *"Neither"* (insufficient evidence to make the comparison). For example, Maria has a higher BDI-II depression score than Priya, thus, Maria > Priya in the comparison. **(3) Ranking Evaluation:** Finally, we compare these comparisons against the ground-truth BDI-II scores. We consider a correct prediction when the LLM acting as a judge chooses the most severe LLM persona from the comparison.

We conducted our experiments including four open-source models that vary in parameter count and training paradigm: llama3-8b-instruct [8], deepseek-r1-14b [9], gemma-3-27b [15] and llama3-70b-instruct [8]. Since each persona is compared to each other in a pairwise approach, the experiment involves a total of 66 comparisons ($\binom{12}{2} = 66$). The results, presented in Table 2, highlight both the general precision of the prediction and key trends regarding incorrect predictions and "Neither" responses.

Overall performance. Among the four judges, the DeepSeek R1-14B variant achieved the best accuracy (86.36%) and never decided on the "Neither" option. Llama 3.1 70B ranked second (81.67%) but declined to decide in six comparisons, indicating a more cautious policy. The two remaining models: Gemma 3.1 27B and Llama 3.1 8B obtained around 77% and 76% accuracy. These results show that model size alone does not guarantee higher accuracy, and that explicit uncertainty is only present in the Llama models.

Incorrect Predictions. All models shared certain challenging cases. For instance, *Marco* (BDI-II=38, severe level) was misclassified a few times, whether compared to Moderate-level personas such as *Linda* (BDI-II=28) or *Laura* (BDI-II=23). This result could indicate that the Marco transcripts, though intended to convey severe symptoms, did not provide enough explicit cues for the model to confidently rank him above someone in the moderate category.

Pair errors with the same vs. different depression levels. Errors where the pairs come from the same level are less concerning from a diagnostic perspective, as individuals within the same BDI-II

Table 3: Mean attribute scores (1–5) for each LLM persona, grouped by depression-severity level.

Severity	Persona	General interaction			Depression-oriented			
		Hum.	Nat.	Flu.	Emo.	Sym.	Eng.	Cog.
Minimal	Maya	3.5	4.0	4.5	4.0	4.0	4.0	3.5
	Noah	3.0	3.0	4.0	3.5	4.0	3.5	4.0
	Priya	3.5	4.5	4.5	4.0	4.5	4.0	4.5
Mild	Alex	3.0	4.0	4.5	4.0	4.0	4.0	4.0
	Ethan	3.5	4.5	4.5	3.5	4.0	4.5	4.5
	Gabriel	3.0	3.5	4.0	4.0	3.5	3.0	3.5
Moderate	James	3.5	4.0	4.5	3.5	4.5	4.0	3.5
	Laura	4.0	4.5	4.5	4.5	4.5	3.5	4.5
	Linda	3.5	4.0	4.5	3.5	3.5	3.5	3.0
Severe	Elena	4.5	4.5	4.5	4.5	4.0	3.5	3.5
	Marco	4.0	4.0	4.5	4.0	4.5	3.5	3.5
	María	4.0	4.0	4.5	3.5	4.0	3.0	3.0

range (e.g., *Alex* at 15 vs. *Gabriel* at 13) can present highly similar clinical profiles. As Table 2 shows, all LLMs make roughly the same number of mistakes when considering the same or different levels of depression. The errors in comparison pairs with different depression levels are low (no higher than 12.95%), so even when the models are wrong, they rarely mix patients whose depression levels are clearly different.

4.2 Clinical Professional Assessment for Simulation Validation

To complement the automatic evaluation of LLM, we recruited two additional certified clinical psychologists (not involved in the simulation design) to rate the 12 LLM personas on a 1–5 Likert scale [10]. We asked them to fill out a form to examine the realism of simulated patients and their behaviours to reflect the severity of the encoded depression and key symptoms. The form covered two dimensions: (*i*) General interaction quality on three attributes: Humanness, Naturalness, Fluency. (*ii*) Depression diagnosis-oriented assessment, with four attributes: Emotional consistency, Symptom realism, Engagement/Responsiveness, Cognitive load & processing style[4]. Each rater conducted conversational interactions with each persona and received the corresponding full interview transcript to complete the form independently. Table 3 reports the mean scores (1-5) per LLM persona for all considered attributes.

Overall evaluation results. Across all personas and attributes, the overall mean is 3.92, indicating that clinicians judged that the simulated patients were well above the midpoint of adequacy. Regarding the two dimensions, the general interaction attributes average 4.01, whereas the four depression-orientated items average 3.84. Thus, while language quality is perceived as strong, symptom portrayal remains slightly lower, suggesting that the pipeline captures psychological cues almost as convincingly as it renders conversational flow.

Differences based on severity levels. Averaging scores within each severity level (Table 3) reveals little differences. The general quality of the interaction increases slightly with severity: from 3.83

---

[4]The guidelines and evaluation instructions are accessible in the repository.

in the minimal group to 4.28 in the severe group. Thus, clinicians perceived the more symptomatic personas equally or even more natural and fluent than the milder ones, suggesting that our prompt design maintained conversational coherence even when simulating severe cases. In contrast, the mean for depression-oriented attributes is better in the minimal and moderate bands (3.96 and 3.83, respectively) and declines to 3.71 for the severe band. This drop is primarily driven by lower scores of participation and cognitive load, which is consistent with the clinical expectation that severely depressed patients participate less and exhibit cognitive slowing [23].

5 Findings, Discussions and Conclusions

In general, our evaluation shows that the TalkDep simulation pipeline produces clinically relevant and conversationally coherent patient personas at a range of depression severity levels. Both LLMs and clinical experts as assessors confirmed that the simulated dialogues reliably reflect the intended BDI-II-based symptom profiles. LLM judges achieved up to 86% precision in pairwise severity comparisons, indicating that the depression signals in our dialogues are detectable and interpretable. Clinical professionals rated personas highly on the dimensions related to interaction and symptoms, with an average score of 3.92 out of 5. People in the severe group were judged fluent and natural in conversations, while lower scores on engagement and cognitive activity matched clinical expectations. However, some difficulty remained in distinguishing adjacent severity levels, especially when moderate personas expressed strong emotional cues. This suggests a need for more controlled prompting to ensure a better representation of symptoms.

TalkDep has already proven useful on the eRisk 2025 pilot track and offers a clinically grounded benchmark to detect depression from conversations [22]. Although current coverage is limited to 12 personas, future extensions could introduce more diverse profiles and modalities. In summary, our findings demonstrate that TalkDep offers a viable resource to benchmark dialogue-based depression detection systems. The evaluated personas also provide a solid foundation to support broader research on patient simulation, particularly for mental health education, where realistic and varied case presentations are essential for preparing clinical practitioners. These personas enables controlled experimentation with symptom severity, multi-system evaluation, and invites further research on the generation and assessment of simulated mental health dialogues.

All stages of profile design and evaluation were performed in consultation with licenced clinicians and no real patient data was used. Because the patients are entirely synthetic, the resource poses no risk of re-identification or psychological harm to actual individuals.

GenAI Usage Disclosure

Generative AI tools were used only for minor editorial tasks such as proofreading, grammar and vocabulary refinement, and general language polishing during manuscript preparation. Every core idea, analysis, experiment, and section of prose was developed and written by the coauthors, with no AI-generated text included.

References

[1] Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical psychological science* 6, 4 (2018), 529–542.

[2] Eliseo Bao, Anxo Pérez, and Javier Parapar. 2024. Explainable depression symptom detection in social media. *Health Information Science and Systems* 12, 1 (2024), 47.

[3] Howard S Barrows. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. *Academic medicine* 68, 6 (1993), 443–51.

[4] Lynn Bickley and Peter G Szilagyi. 2012. *Bates' guide to physical examination and history-taking.* Lippincott Williams & Wilkins.

[5] Joshua EJ Buckman, Rob Saunders, Joshua Stott, L-L Arundell, Ciarán O'Driscoll, Molly R Davies, Thalia C Eley, Steven D Hollon, Tony Kendrick, Gareth Ambler, et al. 2021. Role of age, gender and marital status in prognosis for adults with depression: An individual patient data meta-analysis. *Epidemiology and psychiatric sciences* 30 (2021), e42.

[6] David JA Dozois, Keith S Dobson, and Jamie L Ahnberg. 1998. A psychometric evaluation of the Beck Depression Inventory–II. *Psychological assessment* 10, 2 (1998), 83.

[7] ML Good. 2003. Patient simulation for training basic and advanced clinical skills. *Medical education* 37 (2003), 14–21.

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[10] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology* 7, 4 (2015), 396.

[11] Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health* 11, 1 (2024), e59479.

[12] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.* 10570–10603.

[13] Bingqing Lu, Lixia Lin, and Xiaojuan Su. 2024. Global burden of depression or depressive symptoms in children and adolescents: A systematic review and meta-analysis. *Journal of affective disorders* (2024).

[14] Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu, and Dilek Hakkani-Tur. 2023. Using In-Context Learning to Improve Dialogue Safety. In *Findings of the Association for Computational Linguistics: EMNLP 2023.* 11882–11910.

[15] Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *CoRR* (2024).

[16] Modhurima Moitra, Damian Santomauro, Pamela Y Collins, Theo Vos, Harvey Whiteford, Shekhar Saxena, and Alize J Ferrari. 2022. The global gap in treatment coverage for major depressive disorder in 84 countries from 2000–2019: a systematic review and Bayesian meta-regression analysis. *PLoS medicine* 19, 2 (2022), e1003901.

[17] Amin Omidvar and Aijun An. 2023. Empowering conversational agents using semantic in-context learning. In *Proceedings of the 18th Workshop on Innovative use of NLP for Building Educational Applications (BEA 2023).* 766–771.

[18] Dimitrios Papanagnou, Matthew R Klein, Xiao Chi Zhang, Kenzie A Cameron, Amanda Doty, Danielle M McCarthy, Kristin L Rising, and David H Salzman. 2021. Developing standardized patient-based cases for communication training: lessons learned from training residents to communicate diagnostic uncertainty. *Advances in Simulation* 6 (2021), 1–11.

[19] Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2022. Overview of eRisk 2022: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022.* 233–256.

[20] Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2023. Overview of eRisk 2023: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023.* 233–256.

[21] Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2024. Overview of eRisk 2024: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 14959)*, Lorraine Goeuriot, Philippe Mulhem, Georges Quénot, Didier Schwab, Giorgio Maria Di Nunzio, Laure Soulier, Petra Galuscáková, Alba García Seco de Herrera, Guglielmo Faggioli, and Nicola Ferro (Eds.). Springer, 73–92. doi:10.1007/978-3-031-71908-0_4

[22] Javier Parapar, Anxo Perez, Xi Wang, and Fabio Crestani. 2025. eRisk 2025: contextual and conversational approaches for depression challenges. In *European Conference on Information Retrieval.* Springer, 416–424.

[23] Wendy J Phillips, Donald W Hine, and Einar B Thorsteinsson. 2010. Implicit cognition and depression: A meta-analysis. *Clinical Psychology Review* 30, 6 (2010), 691–709.

[24] Misha Sadeghi, Bernhard Egger, Reza Agahi, Robert Richer, Klara Capito, Lydia Helene Rupp, Lena Schindler-Gmelch, Matthias Berking, and Bjoern M Eskofier. 2023. Exploring the capabilities of a language model-only approach for depression detection in text data. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI).* 1–5.

[25] Daun Shin, Hyoseung Kim, Seunghwan Lee, Younhee Cho, and Whanbo Jung. 2024. Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: Instrument validation study. *Journal of Medical Internet Research* 26 (2024), e54617.

[26] Daria Smirnova, Paul Cumming, Elena Sloeva, Natalia Kuvshinova, Dmitry Romanov, and Gennadii Nosachev. 2018. Language patterns discriminate mild depression from normal sadness and euthymic state. *Frontiers in psychiatry* 9 (2018), 105.

[27] Julio C Tolentino and Sergio L Schmidt. 2018. DSM-5 criteria and depression severity: implications for clinical practice. *Frontiers in psychiatry* 9 (2018), 450.

[28] Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024. PATIENT-ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.* 12772–12797.