Resource-Efficient LLMs for Depression Symptoms Screening: Performance and Limitations in Zero Shot Setting

Muhammad Rizwan, Jure Demšar
Faculty of Computer and Information Science
University of Ljubljana, Večna pot 113, Ljubljana 1000, Slovenia
{muhammad.rizwan, jure.demsar}@fri.uni-lj.si

Abstract

Depression is the leading cause of global disability and early detection is crucial for effective intervention. Recent advances in large language models (LLMs) offer potential for analyzing text to identify depression symptoms. This work investigates the zero-shot capability of LLMs to recognize nine DSM5 depression symptoms from short-text inputs. We evaluated eight open LLMs with model sizes ranging from 1.5B to 14B parameters using a clinically annotated dataset and assessed both overall agreement and symptom-level performance. Results indicate that while smaller models exhibit limited clinical accuracy, the Qwen 2.5-7B model achieves substantial performance with a Cohen's Kappa of 0.603 and a Macro F1 score of 0.648. Notably, a performance plateau between the 7B and 14B Qwen variants suggests that model scaling alone does not guarantee improved symptom-level classification, establishing Qwen 2.5-7B as a resource-efficient model. Further analysis of the best-performing model revealed strengths in identifying salient symptoms like suicidal thoughts, but limitations in recognizing core symptoms such as depressed mood and anhedonia. Misclassification analysis reveals that the model frequently misclassifies posts expressing 'depressed mood' as 'no symptom' or vice versa, often overlooking indicators of irritability or social withdrawal. These findings suggest that resource-efficient LLMs can support preliminary symptom screening in zero shot settings, but there is risk of overlooking clinically important symptoms without fine-tuning.

Keywords: DSM-5, Depression Symtpoms, Large Language Models, Mental Health

1. Introduction

According to the World Health Organization (WHO), depression is one of the leading causes of global disability, affecting an estimated 5.7% of the adult population worldwide. Furthermore, it contributes to approximately 727,000 suicide deaths annually[1]. In Europe, mental health conditions, including chronic depression, affect about 7% of the population. In response to this growing burden, 31 countries in the WHO European Region have committed to integrating mental health into all areas of public policy and prioritizing it within national health agendas[2].

Despite the urgent need for large-scale mental health assessment, traditional clinical evaluation remains resource-intensive and difficult to scale. In recent years, social media platforms have emerged as valuable sources of user-generated content that reflects individuals' emotional states and lived experiences. This development has motivated a growing body of research in natural language processing (NLP) aimed at detecting mental health signals from textual data.

Early computational studies demonstrated correlations between linguistic features and depression-related behaviors on platforms such as Twitter and Reddit, using lexicons, topic models, and traditional machine learning classifiers (Liu et al., 2022; De Choudhury et al., 2013; Coppersmith et al., 2014). However, much of this work framed depression detection as a binary or multi-class diagnosis prediction task, frequently relying on self-disclosed diagnoses as ground truth labels. Clinical researchers have criticized such labeling strategies for their noise, demographic bias, and limited clinical validity (Ernala et al., 2019).

To address these concerns, more recent research has shifted from disorder-level classification to symptom-level modeling. Instead of predicting a diagnosis, these approaches aim to detect individual psychological or behavioral indicators corresponding to specific diagnostic criteria. Studies have explored mapping social media language to DSM-5 depression symptoms such as anhedonia, sleep disturbances, and feelings of worthlessness (Manikonda and De Choudhury, 2017; Chancellor et al., 2019). At the same time, there is an increasing trend toward constructing datasets annotated by domain experts and developing standardized evaluation protocols for depression and suicide risk detection (Zhang et al., 2021; Coppersmith et al., 2014). These developments provide an opportunity to evaluate modern NLP systems against clinically

[1] https://www.who.int/en/news-room/fact-sheets/detail/depression

[2] https://www.who.int/europe/news/item/16-06-2025-with-17–of-people-in-the-region-living-with-a-mental-health-condition–31-countries-commit-to-integrating-mental-health-into-all-policies

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 74–81
12 May 2026.

grounded, human-annotated benchmarks.

The emergence of large language models (LLMs) has introduced new possibilities for mental health text analysis. Instruction-tuned and conversational LLMs demonstrate strong generalization capabilities across diverse tasks without task-specific training, enabling zero-shot and few-shot learning paradigms. Recent studies have explored their application to mental health-related tasks, including depression detection, suicide risk assessment, and emotional support generation (Yang et al., 2023; Lan et al., 2025; Jin et al., 2025; Omar et al., 2024). Zero-shot classification using natural language prompts has become a prominent approach for evaluating LLM generalization, where task labels are framed as natural language descriptions to leverage pretrained knowledge without explicit supervision (Zhao et al., 2023; Kojima et al., 2022). In mental health contexts, prompt-based methods offer the advantage of explicitly incorporating clinical definitions, potentially improving alignment with expert annotations.

Nevertheless, the reliability of LLMs for clinically grounded, symptom-level mental health analysis remains insufficiently studied. In particular, relatively few works systematically examine zero-shot performance of resource-efficient, open LLMs for detecting DSM-5 depression symptoms. The extent to which such models can replicate expert-level symptom identification without fine-tuning remains unclear.

In this work, we investigate whether moderately sized, general-purpose open LLMs can identify depression-related symptoms in short texts under zero-shot conditions. We evaluate their agreement with DSM-5 symptom annotations provided by licensed psychologists, focusing particularly on clinically critical symptoms such as suicidal thoughts, worthlessness, and anhedonia. Through a controlled zero-shot evaluation, we aim to assess both the feasibility and the limitations of LLM-based symptom recognition systems and to identify common misclassification patterns across DSM-5 symptom categories[3].

2. Methods

This section details the methodology employed in our study. We first describe the preparation of the ReDSM5 dataset used for training and evaluation. Then, we describe the experimental setup, including the LLMs used in experiments and the prompt-based inference procedure used to assess their zero-shot depression symptom classification capabilities.

[3]Code: https://github.com/rizwan2phd/zeroshot-depression-symptoms-screening-llms

Symptom	Sentence Count
Depressed mood	326
Worthlessness	284
Suicidal thoughts	175
Fatigue	111
Anhedonia	106
Sleep issues	104
Cognitive issues	53
Appetite change	45
Psychomotor	32
None (control)	374

Table 1: **Distribution of DSM-5 Depression Symptoms in the ReDSM5 Dataset**. This table displays the number of sentences labeled by a licensed psychologist as containing each of the listed DSM-5 depression symptoms, alongside a 'None' control class.

2.1. Dataset Preparation

In this study we used the ReDSM5 dataset (Bao et al., 2025), a clinically labeled Reddit corpus curated according to DSM5 depression standards (Tolentino and Schmidt, 2018), (Information Retrieval Lab, University of A Coruña, 2025). The dataset contains Reddit sentences that were reconstructed from an earlier paragraph-level corpus and re-annotated by a licensed clinical psychologist. Every sentence is labeled for the presence or absence of the nine DSM5 depression symptoms: depressed mood, worthlessness, suicidal thoughts, anhedonia, fatigue, sleep issues, cognitive issues, appetite change and psychomotor. In addition to binary symptom labels, the annotator also provided clinical rationales.

The instances with multiple symptom labels represent a relatively small proportion of the overall dataset. To simplify the task for our model and facilitate clear interpretation of results, we removed these multi-label instances, reducing the dataset size by 129 instances. The dataset also contains sentences with the absence of symptoms, we used those as our control class (symptom = none).

This resulted in a refined dataset of 1,610 single label instances encompassing ten unique categories within the DSM5 symptoms, nine original symptom labels and a new `None` category representing the absence of depressive symptoms (control class). The final distribution of the data can be seen in Table 1.

2.2. Experiment Setup

To evaluate the zero-shot depression symptom classification capabilities of LLMs, we employed a fixed prompt-based inference setup using the eight open models with sizes ranging from 1.5B to 14B param-

eters and model families (`Llama`, `Mistral`, `Qwen`). List of all models with parameter size and corresponding performance can be seen in Table 2. The Qwen family (1.5B, 3B, 7B and 14B parameters) is composed of models trained on large scale multilingual corpora with strong coverage of English and Chinese language. These models are instruction tuned to improve reasoning. The latest LLaMA 3.2 family (1B and 3B parameters) represent compact and efficient instruction tuned models, while the Mistral family (7B and 12B parameters) represent instruction tuned models employ optimized Transformer architectures. Details of the models evaluated and their performance are presented in Table 2.

Through the prompt, we asked the model to perform the task of an expert psychologist and label the input text exactly one of the DSM5 major depressive disorder symptoms i.e. depressed mood, worthlessness, suicidal thoughts, anhedonia, fatigue, sleep issues, cognitive issues, appetite change and psychomotor or assign the none label if the text does not fit any of these labels. The exact prompt was as follows:

> You are an expert clinical psychologist trained in DSM5 diagnostic criteria for major depressive disorder. Analyze the text carefully for indicators of depressive symptoms. If the text describes a depressive symptom, classify it into one category. If the text shows NO depressive symptoms (e.g., normal mood, neutral content, unrelated topics), classify it as NONE.
> TASK: Classify the following text into EXACTLY ONE category from the list below. Categories: DEPRESSED MOOD, WORTHLESSNESS, ANHEDONIA, SUICIDAL THOUGHTS, APPETITE CHANGE, SLEEP ISSUES, FATIGUE, COGNITIVE ISSUES, PSYCHOMOTOR.
> Output ONLY the category label, no explanations, no punctuation, no additional text. Choose NONE if the text shows NO depressive symptoms. Your response must be a single word matching one category exactly.

We constrained the LLM output to return only the category label to ensure consistency and eliminate post-processing ambiguity. The `max_tokens` parameter was set to 15 to further enforce single-label output. Any response failing to produce an exact category label from the predefined list, or containing extraneous text, was considered an invalid prediction and excluded from the evaluation. The last column of Table 2 shows the number of valid predictions per model. All experiments were conducted in a zero-shot setting, meaning no task-specific fine-tuning or in-context examples were provided. To ensure deterministic outputs and reproducibility, the temperature parameter was fixed at 0 for all model inferences.

This setup allows us to assess whether computationally efficient models can effectively generalize to depression symptoms identification under strict zero-shot condition.

3. Results and Discussion

This section discusses the performance of LLMs for zero-shot classification of DSM5 depression symptoms. We first present a comprehensive analysis of model agreement and overall performance, and then analyze the symptom-level insights and misclassification patterns to identify both strengths and limitations of our best performing model.

3.1. Human Agreement and Performance Analysis of DSM5 Depression Symptoms

Our zero-shot evaluation of the models for DSM5 depression symptom classification shows clear performance differences across model sizes and architectures. Models scores can be seen in Table 2 and visualized in Figures **??**. Given the class imbalance in the dataset, we rely on Cohen's Kappa (McHugh, 2012) and Macro F1 as evaluation metrics.

The Macro F1 score is calculated as:

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i \tag{1}$$

where N is the number of classes and F1_i is the F1-score for class i. The class-wise F1-score is calculated as:

$$\text{F1}_i = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{2}$$

Thus, the Macro-F1 score in Equation 1 is computed as the average of the class-wise F1-scores defined in Equation 2.

Cohen's Kappa (κ) is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3}$$

where p_o is the observed agreement and p_e is the expected agreement by chance, as shown in Equation 3.

Overall, extremely small models struggle to perform meaningful clinical agreement in a zero-shot setting. *Llama-3.2-1B-Instruct*, for instance, exhibits small agreement (Cohen's $\kappa = 0.060$) and a very low Macro F1 score (0.089). This indicates that models at this scale are incapable of capturing DSM5 symptom distinctions without additional

Model	Parameters	Cohen Kappa	Macro F1	Valid Predictions
Llama-3.2-1B-Instruct	1B	0.060	0.089	1608
Llama-3.2-3B-Instruct	3B	0.412	0.410	1603
Mistral-7B-Instruct-v0.3	7B	0.511	0.587	1523
Mistral-Nemo-Instruct-2407	12B	0.550	0.607	1601
Qwen2.5-1.5B-Instruct	1.5B	0.343	0.349	1572
Qwen2.5-3B-Instruct	3B	0.491	0.532	1602
Qwen2.5-7B-Instruct	**7B**	**0.603**	**0.648**	1605
Qwen2.5-14B-Instruct	14B	0.600	0.651	1537

Table 2: **Zero-Shot Classification Results of LLMs for DSM-5 Depression Symptoms**. Performance metrics (Cohen's Kappa, Macro F1) for several LLMs are shown, evaluated on the ReDSM5 dataset in zero shot settings. This highlights the potential for direct application of LLMs to mental health assessment. Valid predictions represent sentences where the LLM correctly outputs the exact symptom category. Qwen 2.5-7B-Instruct achieved the best overall performance.

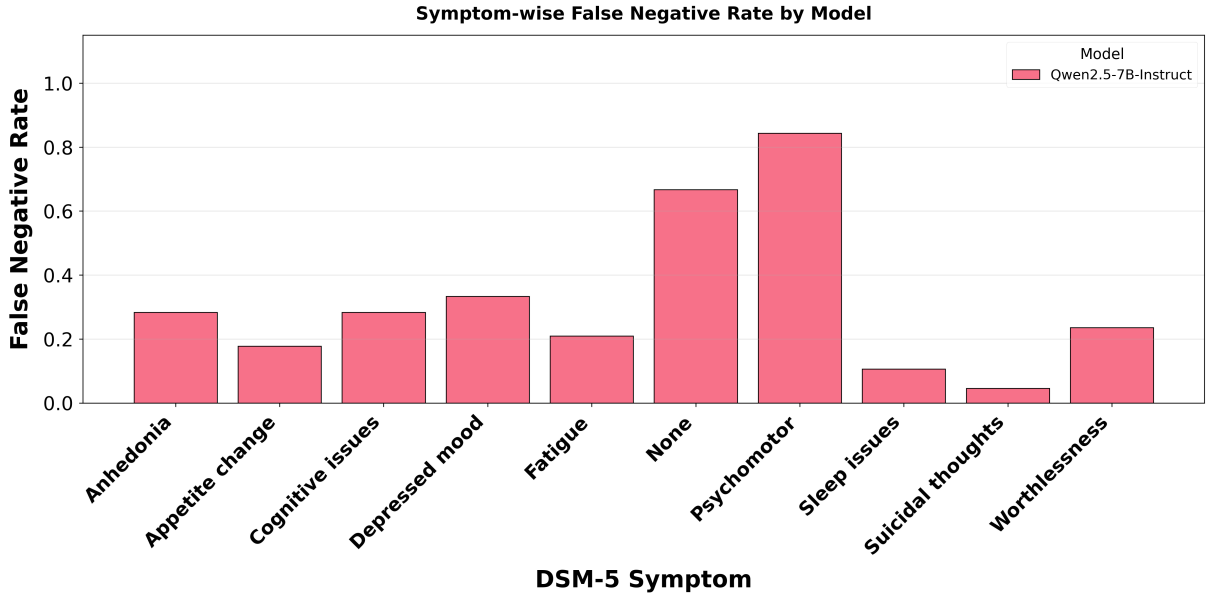

Figure 1: **Symptom-wise false negative rate of the Qwen2.5-7B-Instruct model.** Higher rates indicate a tendency to miss identifying specific symptoms.

human guidance. Performance improves substantially as model size increases into the 3B–7B range, which is the minimum threshold necessary for reliable zero-shot symptom classification. In particular, the *Qwen 2.5* series consistently outperforms size-matched alternatives, which means Qwen architecture and training better suited for depression symptom classification. Among all evaluated models, *Qwen 2.5-7B* emerges as the strongest. It achieves the highest overall agreement with DSM5 labels, with a fair Cohen's κ of 0.603 and a Macro F1 score of 0.648, marginally surpassing both the larger *Qwen 2.5-14B* and *Mistral-Nemo-12B* models. It shows that *Qwen 2.5-7B* offers superior parameter efficiency and more effective zero-shot clinical reasoning.

Interestingly, the performance plateau observed between the 7B and 14B Qwen variants implies that scaling alone does not improve symptom-level classification in the absence of fine-tuning or few-shot prompting. This shows that representational quality of model is more influential than only model size for the zero-shot settings. Based on these findings, we select *Qwen 2.5-7B* as the best performing model for symptom level analyses in upcoming sections.

3.2. Symptom Level Performance

As shown in Figure 3, there is a clear performance plateau between Qwen 2.5-7B and Qwen 2.5-14B, with nearly identical F1 scores across most symptom categories. Although the 14B model demonstrates marginal improvements on a few symptoms (e.g., anhedonia and appetite change), these gains are small and inconsistent. Given this, the

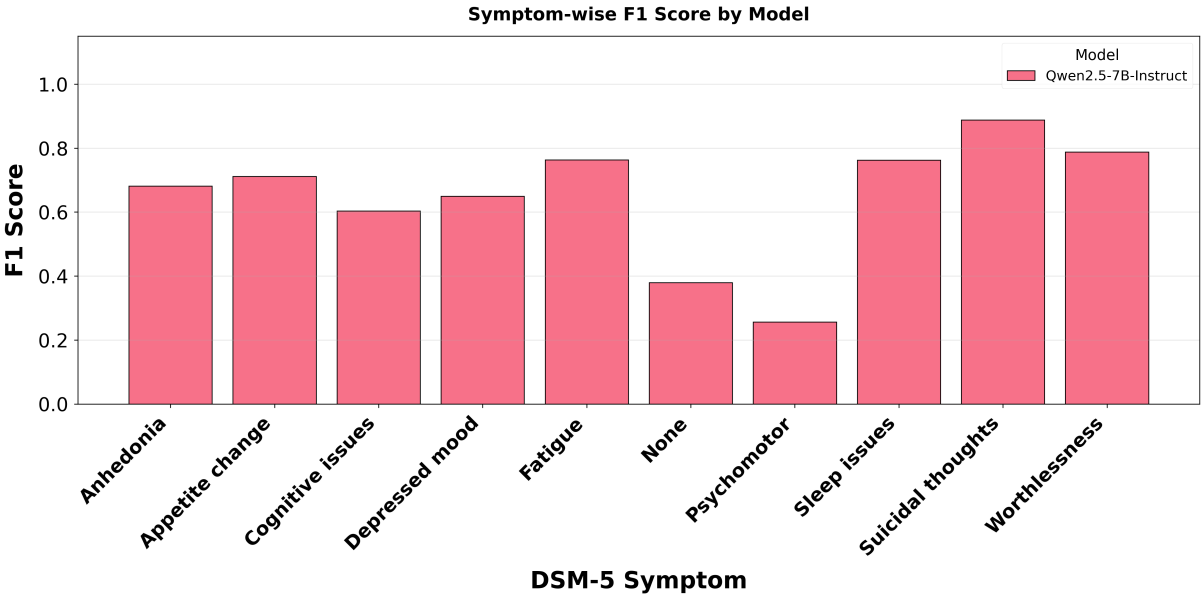

Figure 2: **Symptom-specific F1 scores for Qwen2.5-7B-Instruct.** The figure presents F1 scores across all DSM-5 symptoms, including the None category, illustrating the model's performance in symptom-level detection under zero-shot settings.

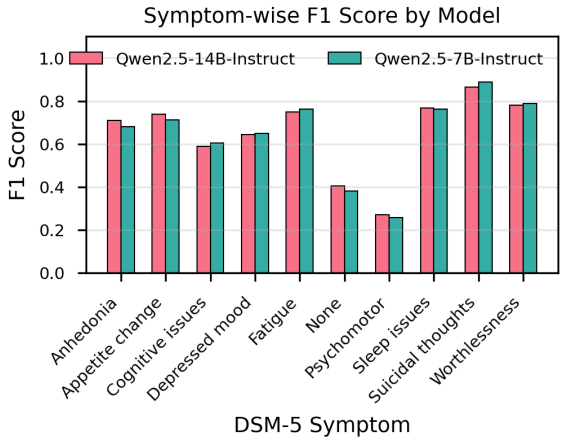

Figure 3: **Symptom-wise F1 comparison.** Comparison of F1 scores across DSM-5 symptoms for *Qwen2.5-7B-Instruct* and *Qwen2.5-14B-Instruct*, showing minimal performance differences between models.

efficiency–performance trade-off favors Qwen 2.5-7B, which achieves comparable results with lower computational cost. Consequently, the 7B model represents a practical choice for deployment and a strong candidate for further fine-tuning to achieve additional performance gains.

The symptom-wise evaluation of our best-performing model, Qwen 2.5-7B, on DSM-5 depression criteria reveals substantial variability across symptoms in terms of F1 score and false-negative rate (FNR). We analyze both the model's strengths and its clinically relevant limitations. The model achieves strong performance on linguistically explicit and high-salience symptoms such as suicidal thoughts (F1 = 0.888), worthlessness (F1 = 0.788), and sleep issues and fatigue (F1 $\approx$ 0.76). Notably, the high recall for suicidal thoughts suggests potential utility in high-sensitivity screening settings. Symptom-wise F1 and FNR scores are shown in Figure 1 and Figure 2, respectively.

In contrast, the model exhibits only moderate performance on core symptoms, including depressed mood (FNR = 0.333) and anhedonia (FNR = 0.283). These elevated false-negative rates are clinically concerning, as prior work (Loas et al., 2018; Auerbach et al., 2022; Gillissie et al., 2023) highlights a strong association between anhedonia and suicidal thoughts. Moreover, suicidal thoughts, depressed mood, and anhedonia are among the most critical symptoms for assessing severe depression risk (Zimmerman et al., 2018). The model performs worst on psychomotor symptoms, with a very high false-negative rate (FNR = 0.884), indicating frequent failure to detect their presence.

Additionally, the *none* category, representing the absence of depressive symptoms, shows poor performance (F1 = 0.379). This indicates a systematic bias toward predicting symptom presence even when no symptoms are present, suggesting that the model may misinterpret neutral or mildly concerning language as clinically relevant, potentially leading to unnecessary alerts in real-world applications.

True Label	Predicted Label	Counts	Sample Sentences (Separated by semi colon)
Depressed Mood	None	72	I know that life is pretty tough and I try to work and safe to have a more secure future; i am angry all the time; Also I have lived recklessly because I thought I didn't had much future; But I never cry , no matter how sad; Where I am neither very happy or sad; I get irritated easily; I'm so irritable all the time.
None	Depressed Mood	66	I feel sad from time to time; I cried so much; Makes me sad ,though; Then I get sad ; I have watched it 3 times and I still cry every time; Iḿ sad and I fear marriage now; I wake up feeling depressed, upset, and anxious; But I never miss my sadness because that's depression and it could get so bad;
None	Sleep Issues	41	Now, i never feel rested in the morning even if i slept 8-9 hours during the night, and it takes a long time to fall asleep ; I usually just get six hours of poor sleep every night; I tend to fell asleep and laid down as soon as I get home; Some nights I would get no sleep at all.

Table 3: : **Misclassification analysis of Qwen 2.5-7B predictions.** The table presents the most frequent misclassifications, including the true label, predicted label, the number of occurrences, and representative example sentences.

3.3. Misclassification Pattern Analysis

In this part of research work, we present results from a detailed exploration of the most common misclassification produced by Qwen 2.5 7B during zero-shot classification of DSM5 depressive symptoms, see Table 3. This is crucial for understanding the limitations of the model when applied to clinical data and for guiding future model improvement.

The most frequent misclassification pattern is to classify depressed mood as none(absence of a symptom). The model struggles to understand that depression does not always present as sadness alone. In several cases, explicit denials of sadness such as *"I wasn't sad or crying anymore"*, *"I don't feel overwhelming sadness"* or *"But I never cry"* lead the model to assume that depression is not present.

Instead of recognizing irritability and anger as possible expressions of a depressed mood, it often treats them as separate, unrelated emotions. Statements such as *"I'm becoming extremely irritable"*, *"constantly mad at little things"* and *"get really irritable"* are frequently misclassified. This suggests a narrow interpretation of depressive symptoms.

Similarly, the model often overlooks signs of social withdrawal. Expressions like *"I had no friends"* or *"Feel alone, that no one has my back"* point to isolation, yet these cues are not consistently recognized.

The second misclassification pattern is the opposite of the first one – classifying sentences without a symptom(the none label) as sentences with a depressed mood symptom. It appears to rely heavily on keywords such as -sad- or -cry- without adequately distinguishing between short-term emotional responses and the persistent low mood. This pattern is especially evident in sentences describing sadness triggered by specific events or ordinary emotional experiences, such as *"I feel sad from time to time"*, *"I cried so much"*, *"Makes me sad though"*, *"Then I get sad"* or *"I have watched it three times and I still cry every time."*

Another prominent pattern among the misclassified instances is the use of quantified sleep-duration constructions without explicit negative qualifiers, typically when explicitly using the keyword sleep along with number of hours. Examples include statements such as *"I normally sleep for 7–8 hours"*, *"I slept 15 hours yesterday"* and *"I sleep 12 hours without waking up"*. Although these sentences describe excessive sleep in context, they are linguistically framed as neutral behavioral reports rather than explicit complaints. The absence of strong distress markers (e.g., insomnia, can't sleep, exhausted) may cause models to interpret them as routine or lifestyle descriptions, leading to misclassification into the control class despite their association with sleep disturbance and depressive symptoms.

4. Conclusion

This study demonstrates the potential of zero-shot, open LLMs for identifying depression symptoms from short text. Our results show notable perfor-

mance differences across model sizes and architectures. We tested several models of different sizes that belongs to different model families. In our case, Qwen 2.5-7B shows the best results for identifying DSM5 depression symptoms. In particular, it achieved fair agreement with DSM5 criteria (Cohen's $\kappa = 0.603$), highlighting its capacity to approximate clinically relevant symptom classification. Furthermore, a symptom-level analysis further revealed that the model performs well in detecting more explicit indicators, such as suicida thoughts, while encountering difficulties to identify depressed mood and anhedonia. Additionally, we observed a systematic bias toward predicting the presence of symptoms even in neutral text, underscoring the need for improved calibration.

Our findings clarify where the model performs reliably and where it falls short, providing direction for targeted refinements. Especially to reduce critical false negatives in high-risk symptom categories. Overall, Qwen 2.5-7B shows promise as a supportive tool for preliminary screening and automated symptom detection. However, we believe that such approaches are not yet ready and should not be used as a substitute for professional clinical assessment.

5. Limitations and Future Work

This study has limitations that should be acknowledged when interpreting the results. First, our evaluation focused on zero-shot performance with a fixed prompt. While this approach was chosen to assess the models' general suitability for initial depressive symptom screening, performance may vary with different prompt formulations. Second, we utilized a single, clinically-annotated dataset (ReDSM5) derived from Reddit posts. This introduces potential bias as the data may not fully represent the broader population of individuals experiencing depression. Furthermore, to simplify analysis, we focused solely on single-label instances, neglecting the common occurrence of co-occurring depressive symptoms represented in the dataset's limited multi-label examples.

Future research address the limitations by exploring few-shot learning approaches and investigating performance on multi-label instances. Specifically, targeted experiments with few-shot learning could reveal whether smaller models within the Qwen family can achieve improved performance in identifying depressive symptoms.

6. Ethics Statement

This study uses the ReDSM5 dataset, which is available under restricted access subject to specific terms and conditions. The data are anonymized and used exclusively for research purposes, with no attempt to identify or re-identify individuals. Given the sensitive nature of mental health content, the proposed study is not intended for clinical diagnosis or deployment without expert supervision.

7. Funding

This publication has received funding from the European Union's Horizon Europe research and innovation program under the Marie Sklodowska-Curie COFUND Postdoctoral Programme grant agreement No.101081355- SMASH and by the Republic of Slovenia and the European Union from the European Regional Development Fund.

8. Disclaimer

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Exacutive Agency. Neither the European Union nor the granting authority can be held responsible for them.

9. References

Randy P Auerbach, David Pagliaccio, and Jaclyn S Kirshenbaum. 2022. Anhedonia and suicide. *Anhedonia: Preclinical, translational, and clinical integration*, pages 443–464.

Eliseo Bao, Anxo Pérez, and Javier Parapar. 2025. Redsm5: A reddit dataset for dsm-5 depression detection. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6323–6327.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology From linguistic signal to clinical reality*, pages 51–60.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–16.

Emily S Gillissie, Gia Han Le, Taeho Greg Rhee, Bing Cao, Joshua D Rosenblat, Rodrigo B Mansur, Roger C Ho, and Roger S McIntyre. 2023. Evaluating anhedonia as a risk factor in suicidality: a meta-analysis. *Journal of psychiatric research*, 158:209–215.

Yu Jin, Jiayi Liu, Pan Li, Baosen Wang, Yangxinyu Yan, Huilin Zhang, Chenhao Ni, Jing Wang, Yi Li, Yajun Bu, et al. 2025. The applications of large language models in mental health: scoping review. *Journal of Medical Internet Research*, 27(1):e69284.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Xiaochong Lan, Zhiguang Han, Yiming Cheng, Li Sheng, Jie Feng, Chen Gao, and Yong Li. 2025. Depression detection on social media with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2155–2171.

Danxia Liu, Xing Lin Feng, Farooq Ahmed, Muhammad Shahid, Jing Guo, et al. 2022. Detecting and measuring depression on social media using a machine learning approach: systematic review. *JMIR Mental Health*, 9(3):e27244.

Gwenolé Loas, Guillaume Lefebvre, Marianne Rotsaert, and Yvon Englert. 2018. Relationships between anhedonia, suicidal ideation and suicide attempts in a large sample of physicians. *PloS one*, 13(3):e0193619.

Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and understanding visual attributes of mental health disclosures in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 170–181.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Mahmud Omar, Shelly Soffer, Alexander W Charney, Isotta Landi, Girish N Nadkarni, and Eyal Klang. 2024. Applications of large language models in psychiatry: a systematic review. *Frontiers in psychiatry*, 15:1422807.

Julio C Tolentino and Sergio L Schmidt. 2018. Dsm-5 criteria and depression severity: implications for clinical practice. *Frontiers in psychiatry*, 9:450.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.

Yipeng Zhang, Hanjia Lyu, Yubao Liu, Xiyang Zhang, Yu Wang, and Jiebo Luo. 2021. Monitoring depression trends on twitter during the covid-19 pandemic: observational study. *JMIR infodemiology*, 1(1):e26769.

Xuandong Zhao, Siqi Ouyang, Zhiguo Yu, Ming Wu, and Lei Li. 2023. Pre-trained language models can be fully zero-shot learners. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 15590–15606.

Mark Zimmerman, Caroline Balling, Iwona Chelminski, and Kristy Dalrymple. 2018. Understanding the severity of depression: which symptoms of depression are the best indicators of depression severity? *Comprehensive psychiatry*, 87:84–88.

10. Language Resource References

Information Retrieval Lab, University of A Coruña. 2025. *ReDSM5: A Reddit Dataset for DSM-5 Depression Detection*. Information Retrieval Lab (IRLab), Universidade da Coruña. Hugging Face Datasets, Mental Health and Clinical NLP Resources, 1.0. PID https://huggingface.co/datasets/irlab-udc/redsm5. Reddit corpus annotated with DSM-5 depressive symptoms. Gated access via Hugging Face. Dataset described in Bao et al., accepted at CIKM 2025.