Evaluating Large Language Models for Depression Symptom Estimation

Dhia Eddine Merzougui$^{1[0009-0003-3676-1469]}$, Gaël Dias$^{1[0000-0002-5840-1603]}$, Jeremie Pantin$^{1[0009-0002-5082-6815]}$, and Fabrice Maurel$^{1[0000-0002-8644-2461]}$

UNICAEN, ENSICAEN, CNRS, GREYC, Normandie Univ, 14000 Caen, France.
{dhia-eddine.merzougui,gael.dias,jeremie.pantin,fabrice.maurel}@unicaen.fr

Abstract. Depression is a mental health disorder that is increasingly prevalent in modern society, impacting individuals' well-being and global public health. Due to its high comorbidity and varied symptom presentation, it poses significant challenges for accurate diagnosis, highlighting the need for advanced tools to assist mental health practitioners in identifying symptoms efficiently. Recent advances in natural language processing have enabled the analysis of text by detecting linguistic patterns associated with depression. This study evaluates the effectiveness of large language models (LLMs) in depression symptom detection, exploring various in-context learning approaches, alongside parameter-efficient fine-tuning techniques for both encoder-based and decoder-based LLMs. By comparing the performance of these methods against existing state-of-the-art approaches, this work reports new state-of-the-art results for depression symptom estimation, and provides insights into the utility of LLMs for improving mental health diagnostics.
The code for our experiments is provided at this anonymous repository.

Keywords: Automated depression level estimation · Mental Health · Natural Language Processing · Large Language Models · In-Context Learning · Parameter-Efficient Fine-Tuning.

1 Introduction

Depression is a major global health concern, affecting millions of individuals and significantly impacting their daily lives. The prevalence of depression is expected to rise due to recent world events [24], further exacerbating its societal and economic burden. Moreover, rising physician burnout rates[1] highlight the need for automated tools to assist in patient care by facilitating early detection, monitoring symptom progression.

Assessing depression is a complex task, with patient-therapist interviews being the standard method used in the medical field to evaluate an individual's mental health. Psychiatrists rely on these conversations to explore various aspects of a patient's life, such as work, living conditions, family dynamics, and relationships, in order to understand their mental state. In addition to these

[1] Medscape study: In 2024, 49% of physicians reported feelings of burnout

interviews, various screening tools have been developed to quantify psychological well-being. One such tool, the Patient Health Questionnaire (PHQ-8) [16], is widely recognized as a valid instrument for diagnosing and measuring the severity of depression. It provides self-reported scores for eight key depressive symptoms: anhedonia, feelings of sadness, sleep disturbances, fatigue, appetite loss, feelings of failure, poor concentration, and physical sluggishness.

In recent years, several studies have explored the topic of automated depression detection using various modalities [10], including natural language processing techniques [17]. These approaches have traditionally framed the task as either a binary classification problem, where individuals are classified as either depressed or not, or as a regression problem, where the PHQ-8 score is automatically learned. But recently, there has been a growing shift towards more nuanced representations of psychiatric syndromes that account for their dimensional and heterogeneous nature. One emerging approach gaining attention is symptom network analysis [4], which implies the analysis of each symptom individually within a graphical dynamic systems. Within that context, Milintsevich et al. [20] first proposed to automatically compute the severity of each symptom instead of a global score for the prediction of depression. However, in terms of computational techniques, previous methods have often been constrained by context length limitations, necessitating hierarchical approaches, and have predominantly relied on encoder-only models.

The advent of large language models (LLMs), including both encoder-based [26] and decoder-based architectures [3, 14], presents new opportunities for improving depression symptom assessment. These models, trained on increasingly large datasets and at greater scales, show promise in aiding mental health diagnosis [15] and have the potential to enhance the accuracy and granularity of symptom classification. In this paper, we explore the application of recent encoder-based and decoder-based LLMs in In-Context Learning (ICL) and Parameter-Efficient Fine-Tuning (PEFT) configurations. Specifically, we investigate zero-shot, few-shot, and Chain-of-Thought (CoT) [27] prompting strategies, as well as Low-Rank Adaptation (LoRA) [12] techniques with both deep and shallow classification heads. Additionally, we examine the potential benefits of reasoning-tuned models for improving depression symptom assessment. Surprisingly, the zero-shot learning strategy reports new state-of-the-art results for depression symptom estimation, and provides new insights into the utility of LLMs for improving mental health diagnostics.

2 Related Work

Automated depression estimation involves using computational models to predict depression severity based on patient-therapist interactions, typically through the analysis of visual, acoustic, and textual features mapped to PHQ-8 scores.

A key research direction is multimodal fusion, where integrating multiple data sources improves predictive accuracy. Qureshi et al. [22] and Ray et al. [23] demonstrate that attention-based fusion networks enhance performance by prioritizing salient features across modalities. Additionally, hierarchical models have been leveraged to encode interview structures, with affective information—such as sentiment and valence—further refining predictions [28].

Multi-task learning has also proven effective, as depression is closely linked to emotion regulation. Qureshi et al. [21] show that jointly estimating depression severity and emotion recognition improves classification and regression outcomes. Demographic-aware modeling has similarly yielded advancements, with gender-sensitive approaches outperforming gender-agnostic models in certain cases [4, 23].

An alternative perspective is provided by Milintsevich et al. [20], who frame depression classification as a symptom profile prediction problem, training a multi-target hierarchical regression model to predict individual depression symptoms from interview transcripts. Agarwal et al. [1] emphasize the importance of discourse structure in mental health assessments, developing multi-view architectures that segment transcripts into sentence-based views, which are processed both independently and in an interconnected manner to capture intra-view and inter-view dependencies. Given the increasing prominence of language models in natural language processing, Ji et al. [13] fine-tune various BERT-based models on mental health datasets, contributing a domain-specific masked language model for mental health text representation. Lau et al. [17] address the scarcity of large-scale, high-quality mental health datasets by advocating for prefix-tuning as a parameter-efficient fine-tuning strategy for language models in this domain.

The remarkable reasoning power of LLMs has fostered continued research in that direction, Chen et al. [5] being a precursor in the field. They propose a structural element graph, which transforms the clinical interview into an expertise-inspired directed acyclic graph for comprehensive modeling. Additionally, they further empower their model by devising principle-guided data augmentation with LLMs to supplement high-quality synthetic data and enable graph contrastive learning. However, the effectiveness of LLMs has yet to be fully evaluated for the task across different training and inference paradigms.

3 Methodology

To study the performance of LLMs in estimating depression symptoms, instead of diagnosing depression as a discrete task, we evaluate different encoder-based and decoder-based models in both ICL and fine-tuning (FT) configurations on the DAIC-WOZ dataset.

3.1 DAIC-WOZ Dataset

The DAIC-WOZ dataset [11] consists of 189 clinical interviews in a dialogue format between a virtual assistant, Ellie, and a human subject. Ellie's responses are selected from predefined prompts, varying between interviews. The dataset is split into 107 training, 35 validation, and 47 test interviews (see Table 1). Each session includes a PHQ-8 assessment, scoring depression symptoms from

Depression severity	**Data split**		
	Train	**Validation**	**Test**
No symptoms [0..4]	47	17	22
Mild [5..9]	29	6	11
Non-depressed Total	**76**	**23**	**33**
Moderate [10..14]	20	5	5
Moderately severe [15..19]	7	6	7
Severe [20..24]	4	1	2
Depressed Total	**31**	**12**	**14**
Total	**107**	**35**	**47**

Table 1: DAIC WOZ Data split based on depression severity

0 to 24. A cutoff of 10 classifies participants as non-depressed (PHQ-8 < 10) or depressed (PHQ-8 ≥ 10), with further classification into five severity levels. The dataset shows class imbalance, particularly at higher PHQ scores. While an extended version exists, this study focuses on the standard version, as the extended dataset lacks the virtual assistant's dialogue, which has been shown to provide valuable information for depression detection [1].

3.2 Large Language Models

In our experiments, we compare two types of LLM architectures: encoder-based and decoder-based language models. Historically, BERT [9] and RoBERTa [18] have been the most widely used models within the encoder family. However, a recently introduced BERT-based model, ModernBERT [26], enhances the original architecture by training on a much larger dataset and incorporating rotary position embeddings, thus enabling it to process significantly longer contexts (8,192 tokens compared to 512 in the original BERT).

For decoder-based architectures, we evaluate Mistral 7B (Mistral-7B-Instruct-v0.3) [14], Llama 3.1 8B (Llama-3.1-8B-Instruct), and Llama 3.2 1B (Llama-3.2-1B), all of which are compact enough for deployment on edge devices, making them accessible for healthcare professionals. We also evaluate state-of-the-art closed-source models Gemini-1.5-Pro [25] and Gemini-2.0-Flash [2] to have a point of reference for the impact of the size of the models and the disparity between openly available models accessible for healthcare professionals to use on-premise and closed-source third party-hosted models not suited for clinical workflows. More recently, novel LLMs leveraging test-time compute have achieved state-of-the-art performance in logic-oriented tasks [8]. In our experiment, we evaluate a distilled version of DeepSeek-R1-8B, which was trained using Llama 3.1 8B and fine-tuned on a synthetic reasoning dataset generated by the original DeepSeek

[2] Gemini 2.0 blogpost.

R1 model. This evaluation aims to investigate the impact of enhanced reasoning capabilities on depression symptom estimation.

3.3 Learning Configurations

In the ICL configuration, we evaluate all models using greedy decoding (temperature set to 0 and beam size of 1) to ensure deterministic outputs. Specifically, in the few-shot configurations, we select a different example from the DAIC WOZ train set for each run and perform the evaluation five times for up to 3 examples to analyze how the choice of example influences model performance. Note that we use the validation set of the DAIC WOZ to optimize the prompts used in this work. The set of constructed prompts is given in section 7.1 in the appendix.

For the FT configuration, we employ both the PEFT technique LoRA [12] and frozen models with learning classification heads. For both cases, we explore various depths for the classification head, ranging from shallow to deep. We hypothesize that, given the reduced number of trainable parameters in LoRA, the small dataset size will not be a limitation, and that a deeper classification head may more effectively capture the complexities of the task [6]. Each model is trained and evaluated five times using different random seeds to assess stability, with the standard deviation reported in the results table. Due to a technical issue, we were unable to train ModernBERT using LoRA; therefore, we opted for full fine-tuning instead, given its relative small size. Hyperparameters used for training/evaluation are detailed in Table 3 in the appendix section.

For evaluation, we use the test set of the DAIC WOZ, applying different metrics for each task: F1-score (micro and macro) for the binary classification and the 5-class depression severity classification tasks, given the unbalanced nature of our dataset. For PHQ-8 score evaluation, we use the standard regression metrics Mean Average Error (MAE) and Root Mean Square Error (RMSE).

4 Results Analysis

4.1 Results for Automatic Diagnosis

Comparing the various training and inference techniques, the Gemini-2.0-Flash, when prompted in a zero-shot configuration, achieves the highest performance in binary depression classification, establishing a new state-of-the-art result with macro and micro F1-scores of 0.84 and 0.85, respectively. It also demonstrates very strong performance in PHQ-score regression, achieving the lowest RMSE score across configurations. Meanwhile, Mistral-7B-Instruct-v0.3 and DeepSeek-R1-8B achieve the highest performance in severity 5-class classification.

Between the two paradigms, ICL and FT, the ICL configuration unexpectedly outperforms FT overall, challenging conventional perspectives in the field [7]. This suggests that for this specific task, explicitly verbalizing task instructions yields better performance than adapting the model weights.

Within the ICL paradigm, increasing model size (e.g., from smaller Mistral 7B and LLaMA 3.1 8B models to Gemini-1.5-Pro) does not lead to improved

	Model	**Binary classif.**		**PHQ regress.**		**5-level classif.**	
		F$_1$-ma	F$_1$-mi	MAE	RMSE	F$_1$-ma	F$_1$-mi
Zero-shot*	DeepSeek-R1-8B	0.62	0.74	**3.17**	4.88	0.34	**0.62**
	Gemini-1.5-pro	0.64	0.74	3.49	4.53	0.27	0.53
	Gemini-2.0-flash	**0.84**	**0.85**	3.47	**4.17**	0.31	0.43
	Llama-3.1-8B	0.69	0.72	4.02	5.19	0.41	0.43
	Mistral-7B	0.78	0.83	3.45	4.73	**0.44**	0.55
Few-shot	DeepSeek-R1-8B-[1S]	0.49$_{(0.05)}$	0.66$_{(0.10)}$	4.43$_{(0.97)}$	5.91$_{(0.71)}$	0.22$_{(0.06)}$	0.46$_{(0.15)}$
	Gemini-1.5-pro-[1S]	0.51$_{(0.04)}$	0.69$_{(0.02)}$	3.66$_{(0.23)}$	5.21$_{(0.32)}$	0.26$_{(0.04)}$	0.56$_{(0.03)}$
	Gemini-2.0-flash-[1S]	**0.75**$_{(0.03)}$	**0.79**$_{(0.03)}$	**3.23**$_{(0.11)}$	**4.22**$_{(0.07)}$	**0.33**$_{(0.07)}$	0.51$_{(0.05)}$
	Llama-3.1-8B[1S]	0.71$_{(0.06)}$	0.77$_{(0.03)}$	3.4$_{(0.33)}$	4.74$_{(0.44)}$	**0.33**$_{(0.06)}$	**0.57**$_{(0.03)}$
	Mistral-7B-[2S]	0.60$_{(0.12)}$	0.74$_{(0.06)}$	3.93$_{(0.61)}$	5.67$_{(0.59)}$	0.30$_{(0.09)}$	0.55$_{(0.06)}$
CoT*	DeepSeek-R1-8B	0.62	0.74	3.79	5.28	0.31	0.57
	Gemini-1.5-pro	0.66	0.74	3.43	4.51	0.26	0.47
	Gemini-2.0-flash	**0.74**	**0.81**	3.19	**4.23**	**0.36**	0.6
	Llama-3.1-8B	0.64	0.74	**2.89**	4.32	0.33	**0.62**
	Mistral-7B	0.7	0.77	3.68	5.0	0.27	0.51
Head-only	DeepSeek-R1-8B-[D]	**0.69**$_{(0.02)}$	**0.79**$_{(0.01)}$	**4.16**$_{(0.28)}$	**5.46**$_{(0.37)}$	0.23$_{(0.02)}$	**0.49**$_{(0.01)}$
	Llama-3.1-8B[D]	0.55$_{(0.09)}$	0.74$_{(0.03)}$	4.55$_{(0.40)}$	5.96$_{(0.56)}$	0.19$_{(0.02)}$	0.46$_{(0.01)}$
	Llama-3.2-1B-[S]	0.41	0.7	5.14$_{(0.09)}$	6.52$_{(0.19)}$	0.18$_{(0.02)}$	0.41$_{(0.04)}$
	Mistral-7B-[S]	**0.71**$_{(0.06)}$	**0.8**$_{(0.03)}$	**3.95**$_{(0.17)}$	**5.04**$_{(0.23)}$	**0.24**$_{(0.03)}$	0.45$_{(0.04)}$
	ModernBERT-[D]	0.34$_{(0.09)}$	0.54$_{(0.20)}$	8.21$_{(3.38)}$	9.53$_{(3.36)}$	0.11$_{(0.04)}$	0.3$_{(0.15)}$
PEFT	DeepSeek-R1-8B-[D]	0.62$_{(0.11)}$	**0.77**$_{(0.04)}$	**4.18**$_{(0.32)}$	**5.48**$_{(0.45)}$	**0.23**$_{(0.04)}$	**0.49**$_{(0.02)}$
	Llama-3.1-8B[S]	0.56$_{(0.08)}$	0.73$_{(0.03)}$	4.3$_{(0.13)}$	5.5$_{(0.21)}$	0.22$_{(0.03)}$	0.46$_{(0.02)}$
	Llama-3.2-1B-[S]	0.41	0.7	5.13$_{(0.09)}$	6.55$_{(0.24)}$	0.17$_{(0.02)}$	0.43$_{(0.04)}$
	Mistral-7B-[S]	**0.64**$_{(0.11)}$	**0.77**$_{(0.05)}$	4.28$_{(0.32)}$	5.62$_{(0.50)}$	0.22$_{(0.02)}$	0.46$_{(0.04)}$
	ModernBERT-[D]	0.34$_{(0.09)}$	0.54$_{(0.20)}$	8.06$_{(3.14)}$	9.3$_{(3.19)}$	0.1$_{(0.05)}$	0.29$_{(0.16)}$
SOTA	Agarwal et al. [2]	0.81$_{(0.01)}$	—	—	—	—	—
	Milintsev. et al. [19]	—	—	3.59$_{(0.31)}$	—	—	—
	Fang et al. (t) [10]	—	—	3.61	4.76	—	—
	Fang et al. (t+v) [10]	—	—	3.36	4.48	—	—
	Chen et al. [5].	0.88**	—	—	—	—	—

Table 2: Evaluation results of overall depression assessment: binary depression classification, PHQ-score regression, and severity 5-class classification. Best result in each section is in bold and best result across all configurations is underlined. [D] stands for deep head and [S] stands for shallow head. [n] indicates the number of examples for few-shot. Note that only best results are reported when different variable configurations are possible.

* For zero-shot and CoT, we use greedy decoding so that the generation is deterministic, therefore the standard deviation is 0.

** Results are not directly comparable since Chen et al. evaluate on the validation set and do not provide results for the test set, although the data is available.

performance. Interestingly, the expected performance gains typically associated with few-shot learning and CoT prompting do not materialize in this task.

With respect to reasoning tuning and by comparing the non-reasoned learning Llama-3.1-8B-Instruct to its reasoned learned version DeepSeek-R1-8B, we observe a performance increase in zero-shot PHQ-score regression and severity 5-class classification. However, the reasoning-tuned model performance degrades under CoT prompting. In contrast, Llama 3.1-8B achieves the lowest MAE across configurations when prompted with the same approach, thus indicating the importance of CoT learning.

Under the FT paradigm, the Mistral-7B model, when paired with a shallow classification head, achieves the highest average performance, interestingly with its parameters frozen rather than using PEFT. In contrast, ModernBERT exhibits the lowest performance for this task.

4.2 Results for Symptom Severity Estimation

In order to better understand the behavior of the symptom-based classification models, we evaluate the learning models from both the ICL and FT configurations, assessing how well their symptom scores align with the ground truth labels. Results are plotted as radar charts in figure 1 for the two best configurations (ICL with Gemini-2.0-Flash and Mistral-7B-Instruct-v0.3), and overall estimation results by symptom for all learning models are given in Table 4 in Appendix.

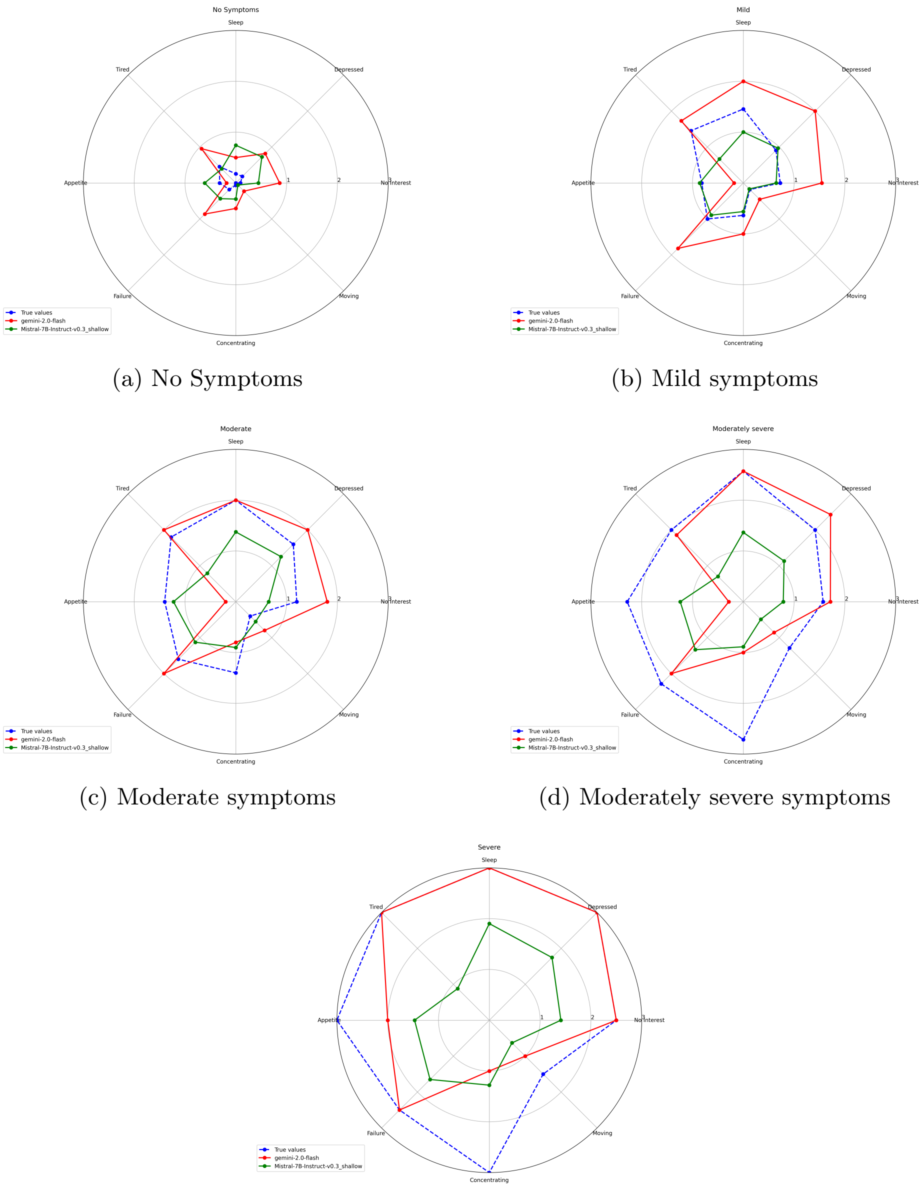
Overall, neither model aligns well with the labels in the no-symptom and mild-symptom categories. However, Gemini-2.0-Flash exhibits a closer alignment with the ground truth labels for moderate to severe symptoms, suggesting improved sensitivity to higher symptom severity levels.

5 Limitations

Although access to the DAIC-WOZ dataset is restricted, we cannot definitively determine whether it was included in the pretraining data of the LLMs used in this study. This potential overlap may influence model performance and limit the generalizability of our findings to real-world applications. Moreover, LLMs commonly exhibit issues related to hallucinations and biases, compromising their reliability for clinical applications. To ensure robust evaluation, future research should validate results across diverse, independent datasets before considering real-world deployment. Additionally, we observed inconsistencies in the performance of the ModernBERT model, as well as instability when using PEFT. Thus, underlying library-related issues may be contributing to its performance variability.

6 Conclusions

This study investigates the application of LLMs for depression symptom detection based on the DAIC-WOZ dataset. In particular, we evaluate recent LLMs



(a) No Symptoms

(b) Mild symptoms

(c) Moderate symptoms

(d) Moderately severe symptoms

(e) Severe symptoms

Fig. 1: Agreement between model-predicted and reported symptom intensity. Gemini-2.0-Flash is shown in red, Mistral-7B-Instruct-v0.3 in green, while the average true symptom value is shown in dotted blue.

of varying sizes, types and architectures using both in-context learning and parameter-efficient fine-tuning techniques on three different tasks, namely binary classification, PHQ-8 score regression and 5-class severity estimation. Overall results show that new state-of-the-art performance can be obtained from zero-shot architectures improving over any other learning strategy. The symptom-based approach allows to verify the accuracy of each model symptom-by-symptom and results acknowledge that different performance values are obtained depending on the patient severity class. Although some limitations exist such as data contamination, the findings of this work demonstrate the potential of LLMs for mental health assessment.

References

1. Agarwal, N., Dias, G., Dollfus, S.: Agent-based splitting of patient-therapist interviews for depression estimation. In: PAI4MH @ NeurIPS (2022)
2. Agarwal, N., Dias, G., Dollfus, S.: Multi-view graph-based interview representation to improve depression level estimation. Brain Informatics **11**(1), 14 (Jun 2024). https://doi.org/10.1186/s40708-024-00227-w, https://doi.org/10.1186/s40708-024-00227-w
3. et al., A.G.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
4. Borsboom, D., Cramer, A.O.: Network analysis: an integrative approach to the structure of psychopathology. Annual review of clinical psychology **9**, 91–121 (2013)
5. Chen, Z., Deng, J., Zhou, J., Wu, J., Qian, T., Huang, M.: Depression detection in clinical interviews with llm-empowered structural element graph. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 8174–8187 (2024)
6. Conneau, A., Schwenk, H., Barrault, L., Lecun, Y.: Very deep convolutional networks for text classification (2017), https://arxiv.org/abs/1606.01781
7. de Andrade, C.M., Belém, F.M., Cunha, W., França, C., Viegas, F., Rocha, L., Gonçalves, M.A.: On the class separability of contextual embeddings representations – or "the classifier does not matter when the (text) representation is so good!". Information Processing Management **60**(4), 103336 (2023). https://doi.org/https://doi.org/10.1016/j.ipm.2023.103336, https://www.sciencedirect.com/science/article/pii/S0306457323000730
8. DeepSeek-AI, et al., D.G.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025), https://arxiv.org/abs/2501.12948
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Fang, M., Peng, S., Liang, Y., Hung, C.C., Liu, S.: A multimodal fusion model with multi-level attention mechanism for depression detection. SSRN Electronic Journal (2022). https://doi.org/10.2139/ssrn.4102839, http://dx.doi.org/10.2139/ssrn.4102839
11. Gratch, J., et al.: The distress analysis interview corpus of human and computer interviews. In: LREC (2014)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685

13. Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., Cambria, E.: MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In: LREC (2022)
14. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), `https://arxiv.org/abs/2310.06825`
15. Kim, J., Leonte, K.G., Chen, M.L., Torous, J.B., Linos, E., Pinto, A., Rodriguez, C.I.: Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. npj Digital Medicine **7**(1), 193 (Jul 2024). `https://doi.org/10.1038/s41746-024-01181-x`, `https://doi.org/10.1038/s41746-024-01181-x`
16. Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B., Berry, J.T., Mokdad, A.H.: The PHQ-8 as a measure of current depression in the general population. Journal of Affective Disorders **114**(1-3), 163–173 (2009)
17. Lau, C., Zhu, X., Chan, W.Y.: Automatic depression severity assessment with deep learning using parameter-efficient tuning. Frontiers in Psychiatry **14** (2023), `https://api.semanticscholar.org/CorpusID:259167991`
18. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019), `https://arxiv.org/abs/1907.11692`
19. Milintsevich, K., Dias, G., Sirts, K.: Evaluating lexicon incorporation for depression symptom estimation (2024), `https://arxiv.org/abs/2404.19359`
20. Milintsevich, K., Sirts, K., Dias, G.: Towards automatic text-based estimation of depression through symptom prediction. Brain Informatics **10**(1), 1–14 (2023)
21. Qureshi, S.A., Dias, G., Hasanuzzaman, M., Saha, S.: Improving depression level estimation by concurrently learning emotion intensity. IEEE Computational Intelligence Magazine **15**(3), 47–59 (2020)
22. Qureshi, S.A., Saha, S., Hasanuzzaman, M., Dias, G.: Multitask representation learning for multimodal estimation of depression level. IEEE Intelligent Systems **34**(5), 45–52 (2019)
23. Ray, A., Kumar, S., Reddy, R., Mukherjee, P., Garg, R.: Multi-level attention network using text, audio and video for depression prediction (2019), `https://arxiv.org/abs/1909.01417`
24. ¸Sim¸sir, Z., c, H.K., Seki, T., Griffiths, M.D.: The relationship between fear of covid-19 and mental health problems: A meta-analysis. Death Studies **46**(3), 515–523 (2022)
25. Team, G., et al., P.G.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024), `https://arxiv.org/abs/2403.05530`
26. Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., Poli, I.: Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference (2024), `https://arxiv.org/abs/2412.13663`
27. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), `https://arxiv.org/abs/2201.11903`
28. Xezonaki, D., Paraskevopoulos, G., Potamianos, A., Narayanan, S.: Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews (2020), `https://arxiv.org/abs/2006.08336`

7 Appendix

7.1 Prompts for the ICL configuration

Zero-Shot prompt

The following is a dialogue between a patient and a therapist. Read it thoroughly and analyze it, trying to determine a score between 0 and 3 for each of the following mental health symptoms. Use the following scale:

- 0: Not present
- 1: Mild (Several days)
- 2: Moderate (More than half the days)
- 3: Severe (Nearly every day)

When scoring, consider both explicit statements and implied behaviors or emotions within the dialogue. If the symptom is not clearly mentioned or implied, assign a score of 0.

Symptoms:

1. LOI: Little interest or pleasure in doing things.
2. DEP: Feeling down, depressed, or hopeless.
3. SLE: Trouble falling asleep, staying asleep, or sleeping too much (sleep disturbances).
4. ENE: Feeling tired or having little energy (fatigue).
5. EAT: Poor appetite or overeating (appetite changes).
6. LSE: Feeling bad about yourself—or that you are a failure or have let yourself or your family down (low self-esteem or guilt).
7. CON: Trouble concentrating on things, such as reading the newspaper or watching television (difficulty concentrating).
8. MOV: Moving or speaking so slowly that other people could have noticed, or being so fidgety or restless that you have been moving around a lot more than usual (psychomotor changes—agitation or retardation).

Make sure to format your answer in the following manner:
LOI - [Number between 0 and 3]
DEP - [Number between 0 and 3]
SLE - [Number between 0 and 3]
ENE - [Number between 0 and 3]
EAT - [Number between 0 and 3]
LSE - [Number between 0 and 3]
CON - [Number between 0 and 3]
MOV - [Number between 0 and 3]

Here is the transcript of the conversation between the patient and the therapist:
{text}

Few-Shot prompt

The following is a dialogue between a patient and a therapist. Read it thoroughly and analyze it, trying to determine a score between 0 and 3 for each of the following mental health symptoms. Use the following scale:
- 0: Not present
- 1: Mild (Several days)
- 2: Moderate (More than half the days)
- 3: Severe (Nearly every day)

When scoring, consider both explicit statements and implied behaviors or emotions within the dialogue. If the symptom is not clearly mentioned or implied, assign a score of 0.

Symptoms:
1. LOI: Little interest or pleasure in doing things.
2. DEP: Feeling down, depressed, or hopeless.
3. SLE: Trouble falling asleep, staying asleep, or sleeping too much (sleep disturbances).
4. ENE: Feeling tired or having little energy (fatigue).
5. EAT: Poor appetite or overeating (appetite changes).
6. LSE: Feeling bad about yourself—or that you are a failure or have let yourself or your family down (low self-esteem or guilt).
7. CON: Trouble concentrating on things, such as reading the newspaper or watching television (difficulty concentrating).
8. MOV: Moving or speaking so slowly that other people could have noticed, or being so fidgety or restless that you have been moving around a lot more than usual (psychomotor changes—agitation or retardation).

Make sure to format your answer in the following manner:
LOI - [Number between 0 and 3]
DEP - [Number between 0 and 3]
SLE - [Number between 0 and 3]
ENE - [Number between 0 and 3]
EAT - [Number between 0 and 3]
LSE - [Number between 0 and 3]
CON - [Number between 0 and 3]
MOV - [Number between 0 and 3]

Here are some examples to help you:
—
{examples}
—
Now, analyze the following dialogue and provide scores for each symptom:
{text}

Chain-of-Thought prompt

The following is a dialogue between a patient and a therapist. Read it thoroughly and analyze it before determining a score between 0 and 3 for each of the following mental health symptoms.

Use the following scale:
- 0: Not present
- 1: Mild (Several days)
- 2: Moderate (More than half the days)
- 3: Severe (Nearly every day)

Instructions for Scoring:
1. Analyze the patient's statements for any explicit or implied mentions of the symptom.
2. Consider frequency and intensity based on the patient's words, tone, and behaviors.
3. Explain your reasoning step by step before deciding on a score.

Symptoms:
1. LOI (Little interest or pleasure in doing things)
2. DEP (Feeling down, depressed, or hopeless)
3. SLE (Sleep disturbances: trouble falling asleep, staying asleep, or sleeping too much)
4. ENE (Feeling tired or having little energy)
5. EAT (Poor appetite or overeating)
6. LSE (Low self-esteem or guilt: feeling bad about yourself or that you are a failure)
7. CON (Difficulty concentrating)
8. MOV (Psychomotor changes: slowed movements or increased restlessness)

Output Format:
For each symptom, provide:
1. A brief analysis explaining how the symptom manifests (or why it does not).
2. A final score in the following format:
LOI - [Number]
DEP - [Number]
SLE - [Number]
ENE - [Number]
EAT - [Number]
LSE - [Number]
CON - [Number]
MOV - [Number]

Here is the transcript of the conversation between the patient and the therapist:
{text}

7.2 Hyperparameters for the FT configuration

Parameter	**Value**
Learning rate	$3e^{-5}$
Epochs	20
Batch size	1
Patience	5
Runs	5
Gradient normalisation	1.0
LoRA target modules	q_proj, v_proj
LoRA rank	8
LoRA alpha	16
LoRA dropout	0.1
Seeds	42, 12345, 9876, 2024, 8675309
Encoder deep cls head	model.hidden_size → 256 → 8
Decoder deep cls head	model.hidden_size → 1024 → 512 → 128 → 8

Table 3: Training Hyperparameters

7.3 Full results for the symptom-based evaluation

	Model	LOI		DEP		SLE		ENE		EAT		LSE		CON		MOV	
		F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE
Zero-shot	DeepSeek-R1-8B	0.362	0.511	0.380	0.617	0.469	0.532	0.365	0.617	0.146	0.936	0.283	0.660	**0.389**	**0.681**	0.253	0.447
	Llama-3.1-8B	**0.452**	**0.426**	0.386	0.894	0.427	0.702	0.496	0.617	**0.200**	**0.809**	0.401	0.830	0.190	0.894	**0.456**	0.340
	Mistral-7B	0.378	0.532	**0.446**	**0.574**	0.308	0.830	**0.596**	**0.426**	0.137	0.915	0.369	0.702	0.294	0.830	0.217	0.340
	Gemini-1.5-pro	0.358	0.574	0.364	0.617	0.493	**0.511**	0.242	0.723	0.187	0.894	**0.511**	**0.553**	0.291	0.787	0.264	**0.319**
	Gemini-2.0-flash	0.228	0.766	0.390	0.745	**0.495**	0.532	0.492	0.553	0.185	0.851	0.493	0.723	0.234	0.830	0.341	0.340
Few-shot	DeepSeek-R1-8B-[1S]	0.252	0.583	0.269	0.719	0.370	0.745	0.245	0.800	0.136	0.962	0.351	0.613	0.181	0.979	0.224	0.383
	Llama-3.1-8B[1S]	0.407	0.460	0.364	0.672	0.474	0.604	**0.549**	**0.506**	**0.190**	**0.838**	0.332	0.664	**0.293**	**0.664**	0.266	0.417
	Mistral-7B-[1S]	0.346	0.511	**0.430**	**0.523**	0.280	0.962	0.399	0.681	0.167	0.860	0.286	0.796	0.274	0.779	0.223	0.383
	Gemini-1.5-pro-[1S]	**0.424**	**0.434**	0.392	0.553	0.471	0.485	0.277	0.694	0.165	0.906	0.323	0.626	0.237	0.843	0.272	**0.298**
	Gemini-2.0-flash-exp-[1S]	0.393	0.553	0.405	0.634	**0.531**	**0.485**	0.508	0.545	0.187	0.847	**0.451**	0.562	0.277	0.753	**0.283**	0.353
CoT	DeepSeek-R1-8B	0.299	0.660	0.388	0.638	0.370	0.638	0.400	0.638	0.145	0.894	0.310	0.702	0.213	0.894	0.213	0.383
	Llama-3.1-8B	0.371	0.489	**0.469**	**0.596**	**0.514**	0.489	**0.599**	**0.426**	0.164	0.894	0.422	0.596	0.272	0.745	0.336	0.319
	Mistral-7B	0.404	0.468	0.391	0.681	0.391	0.766	0.342	0.745	**0.349**	**0.809**	0.399	0.723	0.183	0.957	**0.432**	0.447
	Gemini-1.5-pro	0.403	0.553	0.369	0.766	0.511	**0.447**	0.298	0.681	0.186	0.894	0.456	**0.553**	**0.306**	**0.702**	0.250	0.362
	Gemini-2.0-flash	**0.539**	**0.468**	0.429	**0.596**	0.504	0.553	0.592	0.489	0.179	0.915	**0.543**	**0.553**	0.263	0.894	0.305	**0.298**
Head-only	DeepSeek-R1-8B-[D]	0.217	**0.486**	0.210	**0.618**	0.191	**0.823**	0.179	**0.749**	0.194	**0.762**	0.199	**0.657**	0.255	**0.756**	0.217	**0.396**
	Llama-3.1-8B[D]	0.196	0.525	0.197	0.643	0.200	0.871	0.189	0.778	0.181	0.805	0.207	0.719	0.221	0.787	0.217	0.400
	Llama-3.2-1B-[S]	0.169	0.626	0.156	0.736	0.202	0.962	0.168	0.796	0.168	0.827	0.179	0.855	0.182	0.903	0.217	0.454
	Mistral-7B-[S]	**0.282**	0.522	**0.277**	0.623	**0.244**	0.889	**0.195**	0.819	**0.212**	0.802	**0.324**	0.665	**0.280**	0.787	**0.238**	0.509
	ModernBERT-large[D]	0.104	2.214	0.103	2.302	0.029	3.066	0.080	2.054	0.072	3.736	0.060	2.165	0.028	4.466	0.061	3.620
PEFT	DeepSeek-R1-8B-[D]	0.204	**0.508**	0.207	**0.631**	0.195	**0.827**	**0.189**	**0.736**	0.194	**0.768**	0.196	**0.657**	0.248	**0.773**	**0.217**	**0.382**
	Llama-3.1-8B[S]	0.190	0.537	**0.232**	0.641	**0.235**	0.874	0.182	0.779	0.192	0.813	0.226	0.745	**0.252**	0.792	**0.217**	0.426
	Llama-3.2-1B-[S]	0.175	0.627	0.168	0.759	0.211	0.981	0.150	0.802	0.150	0.828	0.174	0.830	0.178	0.895	**0.217**	0.409
	Mistral-7B-[S]	**0.208**	0.608	0.214	0.684	0.232	0.897	0.174	0.853	**0.202**	0.848	**0.262**	0.699	**0.262**	0.800	**0.217**	0.505
	ModernBERT-large-[D]	0.121	2.116	0.113	2.294	0.029	3.057	0.082	2.145	0.072	3.699	0.060	2.137	0.027	4.474	0.048	3.696
	Milintsevich et al. [19]	–	**0.440**	–	**0.550**	–	**0.630**	–	**0.720**	–	0.690	–	**0.670**	–	**0.670**	–	**0.300**

Table 4: Comparison of all symptoms (F1-Macro and MAE) across sections and models. Best scores within each section are shown in **bold** and the overall best (column-wise) are additionally underlined.