Evaluating Large Language Models for Depression Symptom Estimation

Dhia Eddine Merzougui✉[0009−0003−3676−1469], Gaël Dias[0000−0002−5840−1603], Jeremie Pantin[0009−0002−5082−6815], and Fabrice Maurel[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 prevalent mental health disorder whose varied and comorbid symptom presentation complicates timely and accurate diagnosis. This study evaluates modern encoder- and decoder-based Large Language Models (LLMs) for automated depression symptom estimation using the DAIC-WOZ dataset. We compare In-Context Learning (ICL) strategies (zero-shot, few-shot, chain-of-thought) against parameter-efficient fine-tuning (PEFT/LoRA) and linear probing techniques. Surprisingly, zero-shot ICL achieves new state-of-the-art results, outperforming fine-tuning approaches and prior benchmarks.

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

1 Introduction

Depression poses a significant health challenge, with its complex symptom presentation hindering accurate diagnosis. Early computational efforts primarily focused on overall severity prediction from clinical interviews [13, 14], lacking granular symptom insight. A key advancement was proposed by Milintsevich et al. [12], who predict individual PHQ-8 symptom scores (0-3), offering a more nuanced assessment aligned with clinical needs. However, initial models faced limitations like restricted context length.

The emergence of Large Language Models (LLMs), including large encoders [15] and decoders [4, 7], offers new avenues for this symptom estimation task due to their enhanced context processing and instruction following. While related work explored domain-specific models [9], specific parameter-efficient fine-tuning (PEFT) methods [10], or structural information [3], a systematic evaluation of modern general-purpose LLMs across diverse learning paradigms for the specific task of symptom-level severity estimation [12] remains absent.

This study addresses this gap by evaluating the effectiveness of various modern LLMs on the DAIC-WOZ dataset [6] for the symptom estimation task. We explore two main paradigms: In-Context Learning (ICL), including zero-shot, few-shot [2], and Chain-of-Thought (CoT) prompting [16], and PEFT using

Low-Rank Adaptation (LoRA) [8], alongside linear probing with various linear depth. We compare these approaches against previous results, evaluating encoder- (ModernBERT-base[1]) and decoder-based (Mistral-7B[2], Llama 3 models[3][4], Gemini-Flash/Pro[5], DeepSeek-R1-8B[6]) LLMs. Performance is analyzed across binary classification, PHQ-8 score regression, 5-class depression severity prediction, and individual symptom estimation. Surprisingly, our findings indicate that zero-shot ICL achieves new state-of-the-art results, offering insights into the utility of LLMs for improving mental health diagnostics.

The code and supplementary materials are available at the following repository for reproducibility: https://github.com/HikariLight/depression_estimation.

2 Methodology and Results

We evaluate LLMs for depression symptom estimation on the standard version of the DAIC-WOZ dataset [6]. The task involves predicting the severity score (0-3) for each of the eight PHQ-8 symptoms based on clinical interview transcripts using its standard train, validation, and test splits.

We assess a range of LLMs selected for diverse characteristics: the encoder-based ModernBERT-base representing an improved legacy architecture; compact, open decoder models suitable for edge deployment like Mistral-7B, Llama 3.1-8B, and Llama 3.2-1B; large state-of-the-art models such as Gemini-Flash/Pro; and the reasoning-focused distilled DeepSeek-R1-8B.

Backbones are evaluated models under two main configurations. Within In-Context Learning, we test zero-shot, few-shot (up to 3 examples), and CoT [16] prompting strategies using deterministic decoding (temperature equals to 0, beam size equals to 1). Prompts are optimized on the validation set. For the Fine-Tuning (FT) configuration, we employ PEFT via LoRA and implement standard fine-tuning of classification heads added to frozen base models (i.e., linear probing), exploring both shallow and deep head architectures.

Models are evaluated on the test set using F1-score (macro/micro) for binary and 5-class depression severity classification, and Mean Average Error (MAE)/Root Mean Square Error (RMSE) for PHQ-8 score regression. Results are averaged over five runs with different random seeds. Detailed prompts, full hyperparameter settings, and comprehensive symptom-level results are provided in the public repository: https://github.com/HikariLight/depression_estimation.

Table 1 presents the main results for binary classification, PHQ-8 regression, and 5-class severity classification. Depression score is obtained by summing all symptoms' individual scores. For binary classification, all individuals with

[1] https://huggingface.co/answerdotai/ModernBERTbase
[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[4] https://huggingface.co/meta-llama/Llama-3.2-1B
[5] https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024
[6] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

	Model	Binary classif.		PHQ regress.		5-class 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]	$0.73_{(0.06)}$	$0.81_{(0.03)}$	$3.87_{(0.60)}$	$5.10_{(0.70)}$	$0.26_{(0.05)}$	$0.53_{(0.05)}$
	Llama-3.1-8B[D]	$0.61_{(0.14)}$	$0.77_{(0.05)}$	$4.23_{(0.7)}$	$5.60_{(0.9)}$	$0.22_{(0.05)}$	$0.48_{(0.02)}$
	Llama-3.2-1B-[S]	$0.51_{(0.1)}$	$0.73_{(0.02)}$	$4.58_{(0.32)}$	$5.97_{(0.45)}$	$0.19_{(0.03)}$	$0.44_{(0.02)}$
	Mistral-7B-[S]	$0.74_{(0.01)}$	$0.81_{(0.01)}$	$3.86_{(0.11)}$	$4.92_{(0.15)}$	$0.24_{(0.03)}$	$0.43_{(0.04)}$
	ModernBERT-[D]	$0.48_{(0.06)}$	$0.70_{(0.01)}$	$5.50_{(0.10)}$	$6.60_{(0.15)}$	$0.14_{(0.02)}$	$0.31_{(0.06)}$
PEFT	DeepSeek-R1-8B-[S]	$0.73_{(0.09)}$	$0.80_{(0.05)}$	$3.84_{(0.23)}$	$5.04_{(0.33)}$	$0.27_{(0.08)}$	$0.50_{(0.04)}$
	Llama-3.1-8B[S]	$0.67_{(0.05)}$	$0.77_{(0.03)}$	$4.0_{(0.09)}$	$5.33_{(0.15)}$	$0.22_{(0.04)}$	$0.46_{(0.03)}$
	Llama-3.2-1B-[S]	$0.61_{(0.11)}$	$0.74_{(0.04)}$	$4.53_{(0.51)}$	$5.73_{(0.68)}$	$0.21_{(0.05)}$	$0.43_{(0.06)}$
	Mistral-7B-[S]	$0.70_{(0.06)}$	$0.79_{(0.03)}$	$3.74_{(0.16)}$	$4.83_{(0.24)}$	$0.25_{(0.01)}$	$0.48_{(0.04)}$
	ModernBERT-[S]	$0.50_{(0.05)}$	$0.68_{(0.02)}$	$5.34_{(0.05)}$	$6.35_{(0.14)}$	$0.15_{(0.03)}$	$0.29_{(0.05)}$
SOTA	Agarwal et al. [1]	$0.81_{(0.01)}$	—	—	—	—	—
	Milintsev. et al. [11]	—	—	$3.59_{(0.31)}$	—	—	—
	Fang et al. (t) [5]	—	—	3.61	4.76	—	—
	Fang et al. (t+v) [5]	—	—	3.36	4.48	—	—
	Chen et al. [3].	0.88**	—	—	—	—	—

Table 1: Evaluation results for depression assessment for binary classification, PHQ-8 regression and 5-class depression-level estimation. Best results per strategy are underlined, while best overall results are in bold. [D] stands for deep head, [S] for shallow head and [n] defines the few-shot example count. Maximum results are shown for variable configurations. Standard deviation results are presented over 5 runs to account for model robustness.

* Zero-shot and CoT use greedy decoding, hence zero standard deviation.

** Chen et al.'s results [3] are on the validation set and as such are not directly comparable.

depression score strictly below 10 are considered non-depressed, the remaining being considered depressed. For 5-class prediction, we use the 5 subclasses presented in [12], and PHQ regression is computed between the summed up depression score with the ground truth.

Overall Performance: Gemini-2.0-Flash using zero-shot ICL establishes new state-of-the-art results for binary classification (F1-macro/micro 0.84/0.85) and achieves lowest RMSE (4.17) in PHQ-8 regression. Notably, zero-shot ICL generally outperforms FT approaches across tasks. Few-shot and CoT prompting do not consistently improve performance over zero-shot ICL. Within FT configurations, Mistral-7B with a shallow head (frozen base model) yields best results among tuned models, while ModernBERT performs poorly.

Symptom-Level Analysis: To understand model behavior at a granular level, we analyze predicted symptom scores against ground truth labels (see Fig. 1[7]). Overall, models struggle to accurately estimate no/mild symptoms, but top performers like zero-shot Gemini-2.0-Flash show better alignment for moderate-to-severe symptom levels, suggesting improved sensitivity to higher severity.



(a) No symptoms (b) Mild (c) Moderate (d) Mod. severe (e) Severe

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

3 Conclusions and Limitations

This study evaluates LLMs for depression symptom estimation on the DAIC-WOZ dataset, comparing In-Context Learning and Fine-Tuning strategies over patient-therapist interviews. Results show that zero-shot ICL yields new state-of-the-art results, surpassing FT approaches and demonstrating the potential of prompting LLMs for nuanced mental health assessment.

However, limitations including potential dataset overlap in LLM pretraining and general model biases require careful consideration and further validation on diverse and private datasets before clinical application. Moreover, LLMs with prohibitively large parameter counts pose practical constraints, as their deployment in real-world clinical settings is hindered by intensive computing demands, lack of privacy safeguards for sensitive medical data, and the non-frugality of such models, making them unsuitable for resource-limited or privacy-critical healthcare environments.

[7] More details are given in the supplementary materials at the following repository: https://github.com/HikariLight/depression_estimation.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agarwal, N., Dias, G., Dollfus, S.: Multi-view Graph-based Interview Representation to Improve Depression Level Estimation
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. 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)
4. DeepSeek-AI, et al., D.G.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025)
5. 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
6. Gratch, J., Artstein, R., Lucas, G.M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., et al.: The distress analysis interview corpus of human and computer interviews. In: LREC. vol. 14, pp. 3123–3128. Reykjavik (2014)
7. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
8. 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)
9. Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., Cambria, E.: MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In: LREC (2022)
10. Lau, C., Zhu, X., Chan, W.Y.: Automatic depression severity assessment with deep learning using parameter-efficient tuning. Frontiers in Psychiatry **14** (2023)
11. Milintsevich, K., Dias, G., Sirts, K.: Evaluating lexicon incorporation for depression symptom estimation (2024)
12. Milintsevich, K., Sirts, K., Dias, G.: Towards automatic text-based estimation of depression through symptom prediction. Brain Informatics **10**(1), 1–14 (2023)
13. 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)
14. Ray, A., Kumar, S., Reddy, R., Mukherjee, P., Garg, R.: Multi-level attention network using text, audio and video for depression prediction (2019)
15. Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., et al.: Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv preprint arXiv:2412.13663 (2024)
16. 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)