

Does Fine-Tuning Hurt Cross-Platform Generalization in Depression Detection?

Stanford CS224N Custom Project

Yanav Lall

Department of Computer Science
Stanford University
yanav@stanford.edu

Abstract

Depression is among the most prevalent mental health conditions worldwide, and detecting depression related language in social media has become an important NLP research area in recent years. However, models trained on one social platform may not generalize well to others because these platforms differ substantially in writing style, vocabulary, and user behaviour. We investigate this by comparing fine-tuned models vs. LLM prompting under in-domain (Reddit data) and cross-domain (mixed platform data) evaluation settings. We observe that fine-tuned models achieve the best in-domain performance but their accuracy drops substantially under cross-platform shifts. Prompting-based methods, on the other hand, are more robust and outperform fine-tuned models on cross-domain data. A training size ablation confirms this pattern – as in-domain training data increases, in-domain performance gets better while cross-domain deteriorates. We also observe that domain-specific pretraining provides only marginal gains and does not eliminate this gap. Overall, our results indicate that fine-tuning on platform specific data induces a tradeoff between in-domain accuracy and cross-platform robustness, while prompting-based methods remain better suited for deployment across social media environments.

1 Key Information

- Mentor: David Anugraha
- External Collaborators: None
- Sharing project: No

2 Introduction

Depression is a leading cause of disability and affects more than 280 million people worldwide [1]. Social media platforms are becoming increasingly popular for people to express emotional distress with some degree of anonymity, thus motivating a growing body of NLP research on depression related language signals. Early NLP work on this task relied on hand-crafted features and traditional classifiers, while more recently transformer models such as BERT [2] fine-tuned on labeled datasets have become the standard approach, reporting strong results on benchmark datasets [3, 4].

However, most of this work trains and evaluates models on data from a single platform – typically Twitter or Reddit. These platforms differ substantially in writing style, vocabulary, and user behaviour. For example - Twitter imposes character limits to encourage brevity, Reddit encourages long-form, while clinical intake forms use formal, structured language. A model trained on Reddit data may pick up surface level cues – like subreddit slang associated with depression or formatting patterns – that do not carry over to other platforms instead of picking up genuine markers of depression. This raises concerns about the reliability of current benchmarks as well as real-world utility of fine-tuned models for this task.

At the same time, large language models (LLMs) have shown strong zero-shot and few-shot results across many NLP tasks with no task specific training, relying on prompting based approaches. These are particularly interesting to clinical NLP since they require no labeled data, can factor in domain knowledge through the prompt design, and unlike fine-tuning, do not directly fit to platform specific training data. This naturally raises the question: for depression-related text classification, does prompting offer a more generalizable alternative to fine-tuning?

We investigated this question through a comparative study of adaptation strategies. Our contributions:

1. Cross-domain evaluation showing that fine-tuning and prompting exhibit different generalization behavior.
2. Analysis of frozen vs. fine-tuned models with general vs. domain-specific pretraining, isolating the role of each factor.
3. A training data ablation showing that generalization gap widens monotonically with more in-domain data.
4. Comparison of four prompting strategies, finding that simpler approaches (zero-shot, few-shot) outperform more structured ones.

3 Related Work

Depression detection from social media. Researchers have been using computational approaches to mental health detection from social media posts for a long time. De Choudhury et al. [5] established the area using linguistic & behavioral features from Twitter. After that, work explored SVMs with LIWC features [6], LSTMs [7], and fine-tuned transformers [3]. Ji et al. [8] released MentalBERT, a BERT model with continued pretraining on Reddit mental health communities. However, most of this work is only evaluating on data from the training platform, hence leaving cross-platform generalization unexamined.

Domain shift and robustness. Domain shift is well-studied in NLP. Gururangan et al. [9] showed that domain-adaptive pretraining can improve in-domain performance but did not extensively study cross-domain performance. Our work measures the specialization-generalization tradeoff specifically for mental health classification.

LLMs for mental health and prompting. Since GPT-3 [10], there has been growing interest in using LLMs without fine-tuning. Wei et al. [11] showed chain-of-thought prompting improves reasoning. Some work has explored LLMs for mental health tasks [12, 13]. However, systematic comparisons between prompting and fine-tuning under a domain shift remain limited.

4 Approach

We frame this as a binary classification problem: given a social media post x , predict label $y \in \{0, 1\}$ with $y = 1$ indicating depression-related content. We compare three families of approaches spanning classical ML, fine-tuned transformer based models, and LLM prompting. In each of these, we evaluate under both *in-domain* settings, where train and test data come from the same platform and *cross-domain* settings, where the model is evaluated on a mixed-platform mental health dataset not seen during training.

4.1 TF-IDF + Logistic Regression

We represent each post x as a TF-IDF vector with up to 50,000 unigram & bigram features and sub-linear tf-scaling. These are then passed to an L2 regularized logistic regression classifier ($C = 1.0$), which minimizes

$$\mathcal{L}_{\text{LR}} = -\frac{1}{N} \sum_{i=1}^N (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)) + C \|\mathbf{w}\|_2^2 \quad (1)$$

4.2 Transformer Models

We used two pretrained encoder-only models for our experiments:

DistilBERT [14]: 6-layer, 66M parameter model which was distilled from BERT-base. Retains 97% of BERT’s performance while being 60% faster, using a knowledge distillation objective during pretraining.

MentalBERT [8]: 12-layer, 110M parameter BERT-base model with further pretrained on Reddit posts from mental health communities. This follows the domain-adaptive pretraining (DAPT) paradigm [9], where a general-purpose model is further pretrained on domain-specific text before fine-tuning.

Both models encode input and pass the [CLS] token representation through a classification head consisting of two layers

$$\hat{y} = \text{softmax}(W_2 \cdot \text{ReLU}(W_1 \cdot \mathbf{h}_{[\text{CLS}]} + b_1) + b_2) \quad (2)$$

Each of the models is evaluated under two settings:

Frozen encoder: All transformer weights are fixed, only the head parameters i.e. W_1, b_1, W_2, b_2 are updated, making it a fixed feature extractor.

Fine-tuned encoder: All weights are updated via backpropagation to allow model to adapt its representations for the task. We optimize using AdamW with learning rate $= 2 \times 10^{-5}$, weight decay $= 0.01$, linear warmup over 10% of steps, and gradient clipping at norm 1.0.

Both settings minimize cross-entropy loss over the training data:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \quad (3)$$

4.3 LLM Prompting

We evaluate four prompting strategies using Claude Sonnet [15], moving from minimal to more structured tasks framing:

Zero-shot: Model receives minimal instruction and the input text. No examples or guidance.

Few-shot: 5 labeled examples are added to the input following the in-context learning paradigm [10].

Chain-of-thought (CoT): Model is told to reason step-by-step about emotional state expressed before classifying [11]. This is to test whether explicit intermediate reasoning improves accuracy.

Clinician persona: Model is instructed to act as a clinical psychologist, evaluating the post for emotional tone, cognitive patterns, and behavioural indicators. This is to test whether domain-specific framing improves classification.

Full prompt templates can be found in the appendix.

5 Experiments

5.1 Data

In-domain: Reddit Depression Dataset [16]. 232,074 posts from r/SuicideWatch and r/Depression (positive) and other subreddits (negative). Labels were derived from subreddit membership, which served as a proxy for depression-related content *rather* than clinical diagnosis. We subsampled 25,000 posts from the full dataset and divided them into 20,000 train, 2,500 validation, and 2,500 test groups i.e. an 80/10/10 split.

Cross-domain: Mental Health Sentiment Dataset [17]. Approximately 51,000 posts from multiple platforms (Reddit, Twitter and others) labeled across 7 mental health categories – Depression, Suicidal, Anxiety, Bipolar, Stress, Personality Disorder, and Normal. We mapped Depression and Suicidal to positive, all others to negative, and sample 2,500 for cross-domain testing (1,231 positive, 1,269 negative). This dataset differs from reddit in sentence length, vocabulary, and style, providing a suitable option for cross-platform evaluation.

Preprocessing of the text required URL removal, stripping Reddit formatting, collapsing whitespace and truncating to 512 words. We also made separate 500-sample subsets for the LLM experiments because of API cost constraints.

5.2 Evaluation

We report precision, recall, F1-score for the positive class, and accuracy. For the generalization analysis, we define and calculate the *generalization gap* as

$$\Delta_{\text{gap}} = \text{F1}_{\text{in-domain}} - \text{F1}_{\text{cross-domain}} \quad (4)$$

and the relative performance drop as $\Delta_{\text{gap}}/\text{F1}_{\text{in-domain}} \times 100\%$.

5.3 Experimental details

All transformer models used a maximum sequence length of 256 tokens; batch size is 32 for DistilBERT and 16 for MentalBERT due to its larger memory footprint. The fine-tuned models trained for 3 epochs, with early stopping based of validation F1-score. The frozen models trained for 5 epochs with a learning rate of 1×10^{-3} . The LLM prompting experiments had maximum output lengths of 50 tokens for zero/few-shot and 500 tokens for CoT and clinical persona so that they could accommodate longer reasoning outputs.

In the training data ablation (Section 6.2), we trained five separate DistilBERT models on subsets of 1K, 2.5K, 5K, 10K, and 20K training examples. All of them were initialized from pretrained weights and were trained with identical hyperparameters.

5.4 Results

5.4.1 Baseline and Transformer Results

Table 1 shows the baseline and transformer results. In-domain, fine-tuned MentalBERT achieves the best F1-score of 0.981 followed by fine-tuned DistilBERT at 0.974. While the frozen models perform well, they are outperformed by the fine-tuned models by close to 4 percentage points.

Cross-domain however, results reverse. Fine-tuned DistilBERT drops from 0.974 to 0.712 which is close to a 27% relative decrease. The TF-IDF baseline (0.782) actually outperforms both fine-tuned transformers in the cross-domain setting which suggests that fine-tuning introduces platform-specific biases that hurt generalization. We also see that the frozen models fare better than their fine-tuned versions in this setting.

Method	Prec	Rec	F1	Acc
<i>In-Domain (Reddit)</i>				
TF-IDF + LogReg	0.939	0.905	0.921	0.922
Frozen DistilBERT	0.921	0.941	0.931	0.930
Fine-tuned DistilBERT	0.979	0.968	0.974	0.974
Frozen MentalBERT	0.938	0.952	0.945	0.945
Fine-tuned MentalBERT	0.985	0.977	0.981	0.981
<i>Cross-Domain (Mixed Platform)</i>				
TF-IDF + LogReg	0.762	0.803	0.782	0.780
Frozen DistilBERT	0.692	0.821	0.751	0.732
Fine-tuned DistilBERT	0.728	0.696	0.712	0.722
Frozen MentalBERT	0.763	0.775	0.769	0.769
Fine-tuned MentalBERT	0.774	0.690	0.729	0.748

Table 1: Baseline and transformer results on in-domain and cross-domain settings.

5.4.2 Prompting Results

Table 2 shows the prompting results. In-domain, the prompting strategies achieve F1 scores between 0.903 and 0.923 which are similar to the TF-IDF baseline and frozen transformers but below the fine-tuned models. Cross-domain, each prompting strategy outperforms each fine-tuned transformer. Few-shot achieves the best cross-domain F1 (0.859), followed by zero-shot (0.854), hence representing a 14-15 point improvement over fine-tuned DistilBERT (0.712).

We notice an interesting pattern in precision-recall tradeoffs. Zero-shot and few-shot prompting show high recall (0.921 – 0.934) but lower precision cross-domain scores, suggesting a bias toward positive predictions. Chain-of-thought shows the opposite – high precision (0.896) but lower recall (0.750), possibly because extended reasoning leads the model to require stronger evidence before classifying a post as depression-related.

Strategy	Prec	Rec	F1	Acc
<i>In-Domain (Reddit), n = 500</i>				
Zero-Shot	0.891	0.957	0.923	0.918
Few-Shot	0.881	0.953	0.916	0.910
Chain-of-Thought	0.937	0.878	0.906	0.907
Clinician Persona	0.884	0.922	0.903	0.898
<i>Cross-Domain (Mixed Platform), n = 500</i>				
Zero-Shot	0.796	0.921	0.854	0.848
Few-Shot	0.795	0.934	0.859	0.852
Chain-of-Thought	0.896	0.750	0.816	0.838
Clinician Persona	0.797	0.847	0.821	0.822

Table 2: Prompting results using Claude Sonnet [15].

5.4.3 Cross-Domain Generalization Summary

Table 3 summarizes our main finding: fine-tuned models drop by 25–27% cross-domain but prompting methods drop by only 6–10%. The best cross-domain method overall (few-shot, F1: 0.859) outperforms the best fine-tuned model (MentalBERT, F1: 0.729) by 13 points.

Method	In F1	Cross F1	Drop (%)
TF-IDF + LogReg	0.921	0.782	15.1
Frozen DistilBERT	0.931	0.751	19.3
Fine-tuned DistilBERT	0.974	0.712	26.9
Frozen MentalBERT	0.945	0.769	18.6
Fine-tuned MentalBERT	0.981	0.729	25.6
Claude (Zero-Shot)	0.923	0.854	7.5
Claude (Few-Shot)	0.916	0.859	6.2
Claude (CoT)	0.906	0.816	9.9
Claude (Clinician)	0.903	0.821	9.0

Table 3: Cross-domain generalization summary. Performance drop = $(F1_{in} - F1_{cross})/F1_{in} \times 100\%$.

6 Analysis

6.1 Isolating Effect of Fine-Tuning

Table 4 shows a controlled comparison isolating the effects of fine-tuning and domain-specific pretraining. For both DistilBERT and MentalBERT, fine-tuning improves in-domain F1 but *decreases* cross-domain F1 by approximately 4 points. This is an interesting finding since the near-symmetry suggests that extra representation capacity that fine-tuning gives to the model goes into learning Reddit-specific lexical and stylistic patterns rather than genuine generalizable markers of depression.

We also observe that domain-specific pretraining i.e. MentalBERT consistently outperforms more-general pretraining of DistilBERT by 1-2 points across all cells, confirming that mental health domain knowledge is useful. But it does not help reduce the generalization gap since we see approximately the same fine-tuning penalty for both models(~ 4 points). This helps us narrow down the reason for cross-domain degradation to be the fine-tuning process rather than insufficient exposure to mental health language during pretraining.

	General (DistilBERT)		Domain (MentalBERT)	
	In F1	Cross F1	In F1	Cross F1
Frozen	0.931	0.751	0.945	0.769
Fine-tuned	0.974	0.712	0.981	0.729
Δ (FT – Frozen)	+0.043	−0.039	+0.036	−0.040

Table 4: Fine-tuning yields symmetric effects: $\sim +4$ points in-domain, ~ -4 points cross-domain.

6.2 Training Data Ablation

Figure 1 shows how performance varies as a function of training set size. In-domain F1 increases from 0.91 at 1K samples to 0.97 at 20K samples while cross-domain F1 decreases from 0.76 to 0.71 over the same range. The generalization gap grows monotonically from 0.15 to 0.26.

This result helps us rule out the explanation that more data would eventually help cross-domain performance. Instead, it suggests that each Reddit-specific training example pushes the model further towards Reddit-specific features. At 1k samples, the model seems to rely mostly on its pretrained representations, achieving nearly equal performance across all domains. But by 20k samples, it has specialized to a point that actively harms transferability. This has a big practical implication: when cross-platform robustness matters, less fine-tuning data may actually be preferable.

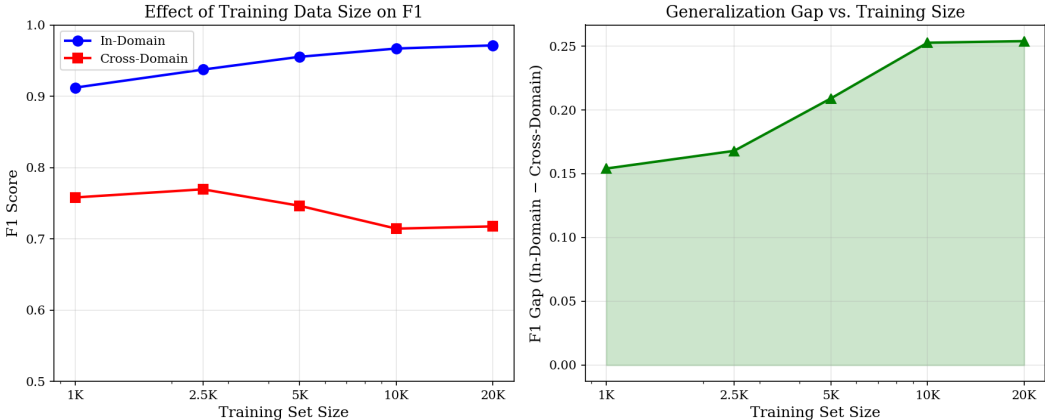


Figure 1: Effect of training data size on in-domain and cross-domain F1 (left) and the resulting generalization gap (right).

6.3 Prompting Strategies

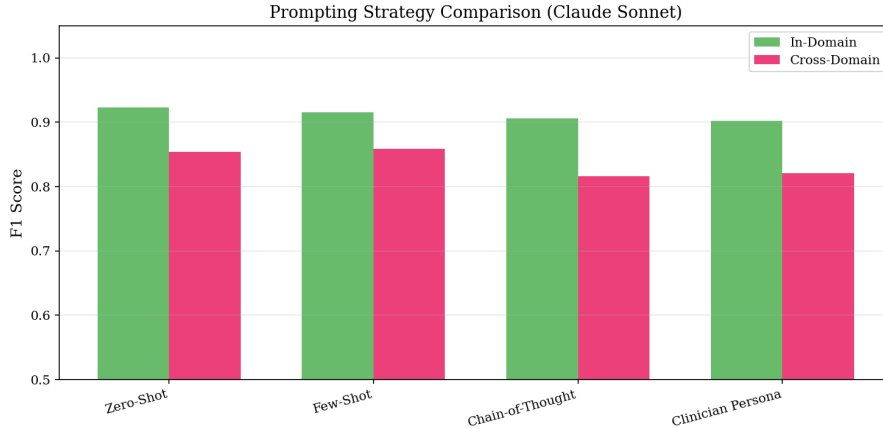


Figure 2: Comparison of four prompting strategies on in-domain and cross-domain data.

Among these prompting strategies (Figure 2), zero and few-shot perform best. Somewhat counter-intuitively, adding examples through few-shot does not improve performance in the in-domain setting but slightly improves cross-domain robustness (+0.005 F1). Chain-of-thought and clinician persona prompting perform worse on both domains. We observed that CoT responses often identified competing emotional responses – say, noting both distress and coping – and struggled to resolve them into a binary label. Zero shot, however, constrained to a single word, was forced to commit. For the task at hand, this constraint seems to help rather than hurt.

6.4 Error Analysis

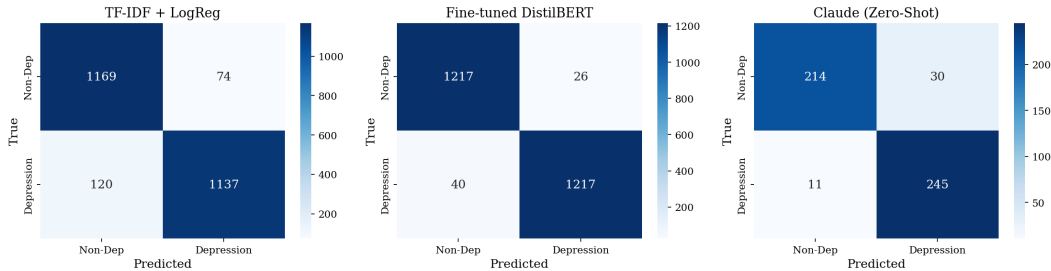


Figure 3: In-domain confusion matrices for TF-IDF + LogReg (left), fine-tuned DistilBERT (center), and Claude zero-shot (right).

In-domain, fine-tuned DistilBERT produced only 66 errors out of 2,500 test samples (26 false positives, 40 false negatives). On examining, we see interpretable failure modes.

False positives (i.e. predicted depression, actually not) often involve posts that use depression-related language in non-depressive contexts. Several involve sarcasm or humor (example - “I just wanna die... noone will care anyway” from a post that was clearly joking about a song lyric). Other ones express mild and temporary negative emotions – “feeling a little down recently” – that the model over-interprets. It seems to hone in on surface-level lexical cues (“die” “feeling down,” “can’t take this”) without fully grasping context or tone.

False negatives (i.e. missed depression) tend to involve indirect expressions of distress. Posts describing bullying and its emotional toll, or posts framed as seeking social connection (“does anyone want to hang out?”) rather than explicitly disclosing depression, were frequently missed. The model

struggles when the depression-related content is embedded in a narrative or request rather than being stated directly.

Cross-domain errors were more revealing. Many cross-domain false positives involved posts about other mental health conditions (bipolar disorder, general therapy discussions) that share vocabularies with depression but describe different conditions. Cross-domain false negatives tended to be shorter, more casual expressions of distress common on non-Reddit platforms – informal language, abbreviations, fragmented writing that differ from the longer, more structured posts typical of Reddit’s mental health communities.

Claude’s zero-shot approach handles these cases more consistently. On the cross-domain set where DistilBERT fails, Claude correctly classifies many posts that use platform-specific language or express distress indirectly hence aligning with our broader finding: the LLM’s general language understanding transfers across platforms in a way that Reddit-specific fine-tuning does not.

6.5 Limitations

Several limitations should be noted. Both datasets use proxy labels rather than clinical diagnoses: the Reddit dataset derives labels from subreddit membership and the cross-domain dataset uses crowd-sourced annotations. Our classification therefore captures self-reported distress signals rather than diagnosable conditions. The cross-domain dataset aggregates multiple platforms, so we cannot disentangle per-platform effects from other distributional differences. Prompting experiments were conducted on 500-sample subsets due to cost, which limits statistical precision. We also evaluate a single LLM (Claude Sonnet) but results may differ for other models or model sizes. And finally, the Reddit dataset’s use of subreddit membership as labels means some positive-class posts may be support messages from non-depressed users, hence introducing label noise.

7 Conclusion

In this report, we present a comparative study of adaptation strategies for depression-related text classification under cross-platform domain shift. Our main finding is that while fine-tuning produces the best in-domain models, it also introduces a specialization-generalization tradeoff that gets worse with more training data and is also not resolved by domain-specific pretraining.

LLM prompting strategies – particularly zero-shot and few-shot – achieve competitive in-domain performance with substantially better cross-domain robustness. Zero-shot prompting outperforms the best fine-tuned model on cross-domain data by over 14 F1 percentage points (0.854 vs. 0.712) while requiring no labeled training data.

These results have implications for the deployment of mental health NLP tools. When a model operates on one, known platform with available labeled data, fine-tuning remains optimal. But when cross-platform robustness is needed – which is often the case for monitoring or screening applications – prompting-based approaches offer a more reliable alternative. Future work could explore hybrid approaches such as prompt-tuning or retrieval-augmented generation that might combine the in-domain strength of fine-tuning with the cross-domain robustness of prompting. Evaluation on a broader range of platforms – including clinical interview transcripts and crisis text lines – would help confirm whether these findings generalize beyond social media. Methods for mitigating platform-specific feature reliance during fine-tuning, such as domain-adversarial training or data augmentation with cross-platform examples, are also promising directions.

References

- [1] World Health Organization. Depressive disorder (depression). *WHO Fact Sheets*, 2023.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.
- [3] Rafał Poświata and Michał Perełkiewicz. OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models. In *Proceedings of*

- the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 276–282, 2022.
- [4] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, pages 31–39, 2015.
 - [5] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 128–137, 2013.
 - [6] David E Losada and Fabio Crestani. A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 28–39, 2016.
 - [7] Ahmed Hussein Orabi, Prasadith Buddhitha, Mahmoud Hussein Orabi, and Diana Inkpen. Deep learning for depression detection of Twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*, pages 88–97, 2018.
 - [8] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mental-BERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of LREC*, 2022.
 - [9] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don’t stop pretraining: Adapt language models to domains and tasks. *Proceedings of ACL*, pages 8342–8360, 2020.
 - [10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
 - [11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
 - [12] Kailai Yang, Shaoxiong Zhang, Bangzhao Zhou, and Erik Cambria. Mentallama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*, 2023.
 - [13] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Yu Hong, et al. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1), 2024.
 - [14] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
 - [15] Anthropic. Claude sonnet model card, 2025.
 - [16] Nikhileswar Komati. Suicide and depression detection dataset, 2021.
 - [17] Suchintika Sarkar. Sentiment analysis for mental health dataset, 2023.

A Appendix

A.1 Prompt Templates

Zero-shot system prompt: “You are a text classifier. You must respond with exactly one word: DEPRESSION or NON-DEPRESSION. No other output.”

Few-shot: Same system prompt, with 5 labeled examples (3 positive, 2 negative) prepended.

Chain-of-thought: “You are a mental health text analyst. When given a social media post, reason step by step about the emotional state expressed, then classify it. End your response with FINAL: DEPRESSION or FINAL: NON-DEPRESSION.”

Clinician persona: “You are a clinical psychologist with expertise in mood disorders conducting a preliminary screening of social media posts. Draw on your clinical training to identify linguistic and emotional markers associated with depressive episodes. End your response with FINAL: DEPRESSION or FINAL: NON-DEPRESSION.”

A.2 Training Data Ablation Details

Train Size	In-Domain F1	Cross F1	Gap
1K	0.912	0.759	0.153
2.5K	0.939	0.771	0.168
5K	0.958	0.749	0.209
10K	0.970	0.721	0.249
20K	0.974	0.712	0.262

Table 5: Full ablation results across training data sizes.

A.3 Data Preprocessing

Text preprocessing: (1) URL removal via regex, (2) removal of Reddit formatting (</r/..., /u/...), (3) whitespace normalization, (4) truncation to 512 words. For the cross-domain dataset, the 7 original categories were mapped to binary labels: “Depression” and “Suicidal” → positive; “Normal,” “Anxiety,” “Stress,” “Bipolar,” “Personality Disorder” → negative. All splits used stratified sampling with seed 42.