Training Compact Models for Mental Wellness Detection: Leveraging LLMs as Teachers

Group 14: Chenchao Lin, Zixuan Wang, Zirong Huang, Xiao Lou, Xiangjie Yuan

1 Introduction

Recognizing and addressing mental health issues is essential for the well-being of individuals. Mental health problems can significantly impact a person's quality of life, leading to emotional distress, impaired functioning, and decreased overall happiness. So, detecting mental health issues early allows for timely intervention and treatment, which can prevent the problem from escalating and becoming more severe. NLP techniques can process and analyze vast amounts of textual data quickly and efficiently. This scalability allows for the examination of large datasets, including social media posts, online forums, chat transcripts, and electronic health records, to identify patterns and trends related to mental health. For our project, the NLP techniques can detect early signs of mental health issues by identifying subtle changes in language use, sentiment, or linguistic patterns.

MultiWD (SATHVIK et al., 2023) and IRF (Garg et al., 2023) are two datasets comprising Reddit posts authored by users, and our objective is to assess the sentiment expressed in each post and identify potential mental health issues in the authors. The LLM analyzes text and generates binary predictions based on its training data, yet training the LLM proves challenging due to resource limitations. The key to the LLM's reliability and accuracy lies in its inferential capabilities, which we aim to transfer to smaller models like Ada, Babbage, and Curie, enhancing their ability to predict mental health concerns. While this doesn't address user privacy concerns entirely, local models such as T5 are tested as well.[1]

2 Methods

Fine-tune-CoT(Ho et al., 2023) Traditionally, a small model will generate answers in a zero-shot way, leading to low answer accuracy. LLMs can generate a thinking process to assist small models instead of letting them answer directly (Wang, L., 2023). This process is called chain of thought (CoT). In order to distill the reasoning prowess of the LLM, we will let LLMs and small models pretend as teachers and students. We need to generate reasoning samples from teacher models and then utilize these to fine-tune the student models. Practically, we will add “think step by step” to the end of each question so that the teacher model will give a thinking process called “rationale.” For the quality of reasoning, we filter out the rationales that lead to wrong answers. Finally, we fine-tune the student models using the rationales that lead to correct answers.

Prompt Engineering In later experiments, rather than simply prompting to "think step by step," we employed a more sophisticated prompt to elicit improved completions from the teacher models for our custom datasets. When we used the original prompt, we observed that the teacher model placed excessive sympathy instead of addressing the underlying issues. After refining the prompt engineering, we observed that the responses were more focused on the question, resulting in more convincing and reasonable answers.

Diverse Reasoning More than traditional fine-tune CoT, we will generate more than one rationale for the same question. From a group of rationales, we will select the ones that lead the

1 Our code implementation and data are available at https://github.com/vinlincc/CSCI-544-Group14-Compact-Models-for-Mental-Wellness-Detection-Leveraging-LLMs-as-Teachers.git.

correct outcomes. And eventually, they will be used to fine-tune the student models.

Auto J (Li et al., 2023) We have proposed that the performances of the student models can be boosted by feeding them rationales in high quality. So, we decided to use Auto-J, also known as Generative Judge for Evaluating Alignment, to combine reasoning prompts and rationales as input to generate a score that evaluates how good the rationales are. The Auto-J will assess the rationale from multiple aspects and give a final score. Auto-J is able to give a rating from 1 to 10, and it selectively filters out all the rationales scored below 6. We expect to enhance the performances of the student models by adopting the remaining ones with high scores.

3 Experiments

Experiment Setup In our experiments, we adopted a multi-faceted approach to fine-tuning and evaluating language models in the context of mental health analysis. The language models used include gpt-3.5-turbo-instruct as the teacher model, Ada, Babbage, and Curie as student models from OpenAI API, t5, and flant5-based local models from hugging face. Local models are fine-tuned with an RTX-4090 GPU.

Dataset and Preprocessing To conduct our experiments, we utilized two key datasets: IRF(Interpersonal Risk Factors) and MultiWD(Multiple Wellness Dimensions). The IRF dataset is meticulously designed to detect interpersonal risk factors, all curated from Reddit posts with human annotations. This dataset is vital for understanding and identifying interpersonal dynamics that may contribute to mental health issues. On the other hand, the MultiWD dataset focuses on detecting multiple wellness dimensions within social media content. MultiWD offers a comprehensive view of various aspects of mental well-being, such as social and emotional health, making it an invaluable resource for analyzing mental health through the lens of social media interactions. The datasets are cleaned so that there are no misleading symbols and notations, and transformed to a format that they formulate binary classification tasks. Each input consists of the Reddit post and the related mental health question, and the expected output is the Chain-Of-Thought (COT) as an explanation and the final result as "yes" or "no."

Prompt Engineering Experiments Initially, we employed a prompt engineering strategy focused on fine-tuning the Chain-of-Thought (CoT) process without incorporating diverse reasoning elements. We compare the zero-shot prompt "let's think step by step" used in the fine-tune-cot paper with our prompt engineer way to prompt the teacher model. Completions generated by both prompts are used to perform fine-tune-cot without diverter reasoning in both OpenAI models. The following experiments are conducted using the engineered prompt.

Baseline and Metrics As a baseline for comparison, we used vanilla fine-tuning methods on OpenAI models and custom models. The discriminant-based models in the mental health domain like mentalBert (Ji et al., 2022) are also investigated as a possible benchmark. However, it turns out that mentalBert is not good enough to be included as a high benchmark that we aim to beat. Our evaluation metrics began with accuracy to gauge the general performance of the models. To obtain a more nuanced understanding, we also employed the F1 score, which provided insights into the precision and recall of the models. We utilized the Auto-J score to evaluate the quality of explanations generated by our models. Because of the limited computational time, the exploration experiments of using Auto-J score as a metric and a filter are limited to fine-tune cot without diverse reasoning to investigate Auto-J usage as a filter to improve the rationales quality.

Auto-J Exploration Experiments In our exploration of the Auto-J scoring system, we employed Auto-J to selectively filter out teacher completions that achieved an Auto-J score higher than 6. This threshold was chosen to ensure that only completions aligning well with human reasoning and preferences were considered for fine-tuning our models. Due to constraints in computational resources and time, we limited this experimental setup to Ada, Babbage, and Curie., all fine-tuned without diverse reasoning. Our objective was to assess and compare the performance of these models in terms of accuracy, F1 score against models fine-tuned through CoT without diverse reasoning and without the Auto-J filter.

Diverse Reasoning Experiments Additionally, we expanded our fine-tuning methods to include diverse reasoning by introducing multiple shots (8 and 16 shots) in the CoT process. Result data,

comparisons with baseline, and discussions are elaborated on in the Results & Discussion section.

Dataset	Ada	Babbage	Curie
Before Prompt Engineering			
IRF	65.38	69.64	69.13
MultiWD	65.07	66.43	69.14
After Prompt Engineering			
IRF	72.34	72.14	71.81
MultiWD	67.38	68.57	68.95

Table 1: **Fine-tune-CoT (Ft-CoT) Performance.** Accuracy (%) of OpenAI models before and after prompt engineering.

Approach	Ada	Babbage	Curie	T5 Base	T5 Small	Flan T5 Base	Flan T5 Small
Accuracy (%)							
Baseline	80.13	79.18	81.74	78.78	76.14	76.65	77.16
Ft-CoT	72.34	72.14	71.81	75.19	72.82	75.59	72.89
Ft-CoT (D = 8)	73.29	74.78	77.15	78.23	75.72	79.51	76.74
Ft-CoT (D =16)	73.29	74.51	N/A	79.38	77.69	79.58	78.70
Ft-CoT (Auto J)	69.23	70.04	71.80	71.53	68.55	N/A	N/A
F1 (%)							
Baseline	84.98	84.98	84.98	84.98	84.98	84.98	84.98
Ft-CoT	81.74	81.74	81.74	81.74	81.74	81.74	81.74
Ft-CoT (D = 8)	82.54	82.54	82.54	82.54	82.54	82.54	82.54
Ft-CoT (D =16)	82.31	82.31	82.31	82.31	82.31	82.31	82.31
Ft-CoT (Auto J)	80.81	80.81	80.81	80.81	80.81	80.81	80.81

Table 2: **Fine-tune-CoT (Ft-CoT) and Baseline Performances.** Accuracy (%) and F1 (%) of OpenAI models and customized models under Fine-tune-CoT (with diverse reasoning) and baseline methods on the IRF dataset. D stands for the degree of diverse reasoning. We could not run a few of the models due to budget and time constraints.

Approach	Ada	Babbage	Curie	T5 Base	T5 Small	Flan T5 Base	Flan T5 Small
Accuracy (%)							
Baseline	84.98	83.92	86.46	84.48	83.23	83.45	83.39
Ft-CoT	81.74	81.56	81.39	82.83	81.43	83.70	81.96
Ft-CoT (D = 8)	82.54	83.40	84.83	84.62	82.87	84.71	83.53
Ft-CoT (D =16)	82.31	83.12	N/A	84.75	84.17	85.00	85.00
Ft-CoT (Auto J)	80.81	80.73	81.85	81.35	80.47	N/A	N/A
F1 (%)							
Baseline	69.92	69.80	69.82	69.84	68.20	68.71	71.18
Ft-CoT	54.96	58.79	60.96	61.83	57.81	63.51	57.22
Ft-CoT (D = 8)	64.48	65.24	66.42	68.30	68.02	68.56	68.75
Ft-CoT (D =16)	66.10	65.79	68.10	69.70	68.79	68.61	68.35
Ft-CoT (Auto J)	53.75	51.47	57.64	57.29	48.49	N/A	N/A

Table 3: **Fine-tune-CoT (Ft-CoT) and Baseline Performances.** Accuracy (%) and F1 (%) of OpenAI models and customized models under Fine-tune-CoT (with diverse reasoning) and baseline methods on the MultiWD dataset. D stands for the degree of diverse reasoning. We could not run a few models due to budget and time constraints.

4 Results & Discussion

Prompt engineering on teacher models has proven to be an effective strategy for enhancing student model performance, particularly for smaller models. As shown in Table 1 above, Ada has experienced a substantial improvement, with accuracy rising by 10% on the IRF dataset and by 3% on the MultiWD dataset after prompt engineering. Similarly, Babbage has demonstrated an approximate 3% increase in accuracy across both datasets. However, larger models such as Curie do not benefit as significantly from prompt engineering, suggesting that this technique is particularly advantageous for smaller models like Ada and Babbage.

Fine-tune-CoT with diverse reasoning approach gradually improves model performances. To examine the learning effects of diverse reasoning and compare it with common Fine-tune-CoT, we applied diverse reasoning with 8 and 16 shots per sample across 7 models on IRF and MultiWD. Table 2 and Table 3 show that diverse reasoning can gradually improve the performance of student models. We noted that smaller models(such like T5 series) using diverse reasoning can beat larger models(such like gpt3 series) that do not use diverse reasoning. With the 8-shot diverse reasoning approach, the T5 Base model attains an F1 score of 68.02 on the MultiWD dataset. This outcome notably exceeds the performance of the Curie model by a margin of 12.06%. This comparison becomes even more pronounced when considering 16-shot, like T5 Base exceeds Curie for 14.34%, further highlighting the superior capabilities of diverse reasoning in handling complex reasoning tasks on small models.

The amount of training data can substantially affect model performances. Tables 2 and 3 show that the Fine-tune-CoT performance of almost all models does not surpass their respective baselines. This trend is due to the size of the dataset diminishing, leading to inadequate training of the models. This phenomenon is especially evident in the Auto J experiments, where only a subset exceeding six scores, as assessed by Auto J, is selected for evaluation from the Fine-tune-CoT datasets. For instance, on the MultiWD dataset, the Ada model achieves a baseline F1 score of 69.92, but this drops to 54.96 in the Fine-tune-CoT approach, indicating a decrease of 21.4%, and further declines to 53.75 in the Auto J setting, marking a reduction of 23.13%. Nevertheless, the implementation of diverse reasoning approach offers a feasible pathway to augment the size of the dataset, thus improving model performances.

5 Conclusion

Our study indicates that the fine-tune Chain-of-Thought (CoT) approach can be adapted to the field of mental health, yet it's crucial to acknowledge that without prompt engineering and diverse reasoning, this approach does not outperform the baseline established by vanilla fine-tuning methods. This limitation highlights the essential role of diverse reasoning in enhancing the performance of the CoT methodology, particularly in specialized areas like mental health. When diverse reasoning is integrated, the CoT approach not only has comparable performance but can surpass baseline, offering benefits such as enhanced privacy, local deployment, and valuable explanations for human reference.

A key insight from our research is the inherent challenge faced by student models in generating answers with CoT as an explanation. This task proves to be more complex compared to vanilla fine-tuning, essentially posing a harder problem for these models. Consequently, the incorporation of diverse reasoning becomes indispensable for achieving performance comparable to baseline models.

As for the integration of Auto-J as a filter module in our fine-tune-CoT pipeline, we observed a performance degradation, likely due to the reduction in the size of the fine-tuning data. Future investigations into the role of Auto-J, particularly when combined with diverse reasoning, are essential to gain a more comprehensive understanding of its efficacy in enhancing data quality and overall model performance.

In conclusion, our study not only contributes valuable insights into the fine-tuning of language models for mental health analysis but also highlights areas for further research. The complex interplay between CoT methodologies, model capabilities, and domain-specific requirements presents both challenges and opportunities for advancing the field of AI in mental health.

6 Individual Contribution

Chenchao Lin: Prompt engineering and generate completions for Fine-tune-CoT and diverse reasoning; local models Fine-tune-CoT and baseline experiments; assist with Auto-J experiments.
Zirong Huang: Reproduction, final baseline finetune, openai model finetune (CoT, 8 shots & 16 shots diverse reasoning).
Zixuan Wang: Preprocess dataset, reproduce experiment results, explore improvement based on the benchmark.
Xiao Lou: Wrote scripts to preprocess Irf and MultiWD datasets; fine-tuned on, inferred completions from, and calculated accuracies of the initial baseline models.
Xiangjie Yuan: Experiment with Custom models for Baseline, Fine-tune-CoT, and Auto J.

References

Garg, M., Shahbandegan, A., Chadha, A., & Mago, (2023). An Annotated Dataset for Explainable Interpersonal Risk Factors of Mental Disturbance in Social Media Posts. arXiv preprint arXiv:2305.18727.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2021). Mentalbert: Publicly available pretrained language models for mental healthcare. arXiv preprint arXiv:2110.15621.

SATHVIK, M., & Garg, M. (2023). MULTIWD: Multiple Wellness Dimensions in Social Media Posts.

Ho, N., Schmid, L., & Yun, S. Y. (2022). Large language models are reasoning teachers. arXiv preprint arXiv:2212.10071.

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K. W., & Lim, E. P. (2023). Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. arXiv preprint arXiv:2305.04091.

Li, J., Sun, S., Yuan, W., Fan, R. Z., Zhao, H., & Liu, P. (2023). Generative judge for evaluating alignment. arXiv preprint arXiv:2310.05470.