TYPE Mini Review
PUBLISHED 12 June 2024
DOI 10.3389/fdgth.2024.1410947

Check for updates

OPEN ACCESS

EDITED BY
Hui Zheng,
Zhejiang Normal University, China

REVIEWED BY
Kausik Basak,
JIS Institute of Advanced Studies and Research, India

*CORRESPONDENCE
Y. H. P. P. Priyadarshana
✉ 2022md05@kuas.ac.jp

RECEIVED 02 April 2024
ACCEPTED 28 May 2024
PUBLISHED 12 June 2024

CITATION
Priyadarshana YHPP, Senanayake A, Liang Z and Piumarta I (2024) Prompt engineering for digital mental health: a short review.
Front. Digit. Health 6:1410947.
doi: 10.3389/fdgth.2024.1410947

# Prompt engineering for digital mental health: a short review

Y. H. P. P. Priyadarshana*, Ashala Senanayake, Zilu Liang and Ian Piumarta

Ubiquitous and Personal Computing Lab, Faculty of Engineering, Kyoto University of Advanced Science (KUAS), Kyoto, Japan

Prompt engineering, the process of arranging input or prompts given to a large language model to guide it in producing desired outputs, is an emerging field of research that shapes how these models understand tasks, process information, and generate responses in a wide range of natural language processing (NLP) applications. Digital mental health, on the other hand, is becoming increasingly important for several reasons including early detection and intervention, and to mitigate limited availability of highly skilled medical staff for clinical diagnosis. This short review outlines the latest advances in prompt engineering in the field of NLP for digital mental health. To our knowledge, this review is the first attempt to discuss the latest prompt engineering types, methods, and tasks that are used in digital mental health applications. We discuss three types of digital mental health tasks: classification, generation, and question answering. To conclude, we discuss the challenges, limitations, ethical considerations, and future directions in prompt engineering for digital mental health. We believe that this short review contributes a useful point of departure for future research in prompt engineering for digital mental health.

KEYWORDS
prompt engineering, digital mental health, natural language processing, large language models, generative artificial intelligence

## 1 Introduction

Even though adapting general-purpose pre-trained large language models (LLMs) to various natural language processing (NLP) tasks such as sentiment analysis has gained a significant attention due to its task-specific fine-tuning capabilities (1), this approach still demands high computational resources and task-specific labelled corpora which make it inappropriate for improving few-shot task performance in complex systems (2). Prompt engineering (PE) has therefore become state-of-the-art (SOTA) for casting various NLP-driven downstream tasks into a general-purpose LLM format (3). As shown in Figure 1, parameter-efficient prompt engineering methods have gained superiority by prepending prompt embeddings to input data while keeping the majority of the LLM frozen (4).

On the other hand, better early identification of human mental disorders has become a vital necessity due to the significant skilled labor requirement for clinical diagnosis-based approaches (5). Even though a few LLM-driven approaches have been introduced for mental disorder detection, fine-tuning their performance is hampered by the limited scalability of the models (6). PE-based methods have recently shown significant improvement for the detection of mental disorders such as depression and anxiety using user-generated text (7).

In this short review, we focused on recently published articles since 2020 by querying four online databases (ACM Digital Library, PubMed, Google Scholar, and IEEE Xplore),

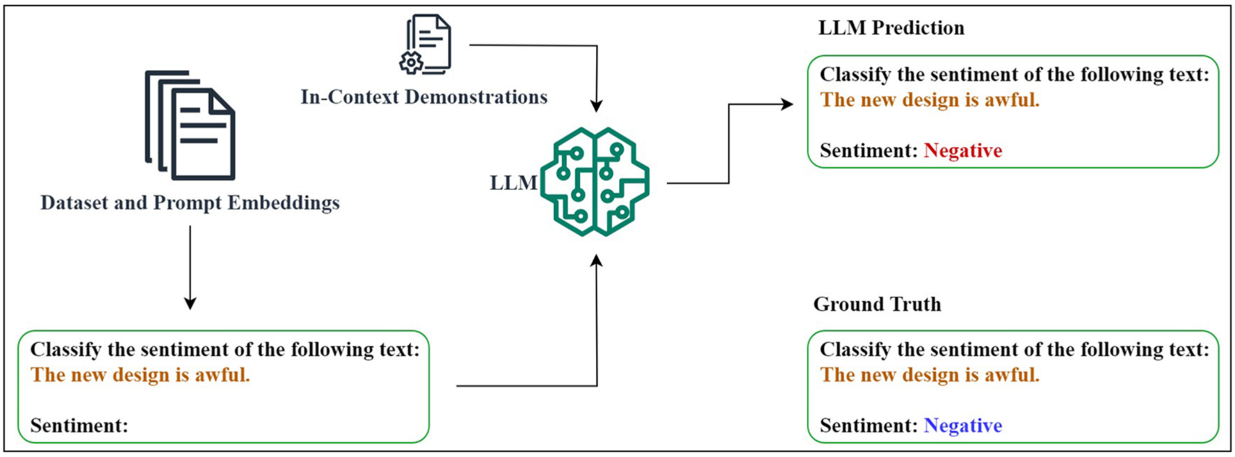
In-Context Demonstrations
Dataset and Prompt Embeddings
LLM
LLM Prediction
Classify the sentiment of the following text:
The new design is awful.
Sentiment: Negative
Classify the sentiment of the following text:
The new design is awful.
Sentiment:
Ground Truth
Classify the sentiment of the following text:
The new design is awful.
Sentiment: Negative

FIGURE 1
Prompt-based LLM fine-tuning for sentiment analysis.

using keywords such as "Prompt Engineering," "Deep Learning for Mental Health," "Deep Learning for Digital Mental Health," "In-context Learning," "Prompt Tuning," "Instruction Prompt Tuning," "In-domain Prompting," "Out-of-domain Prompting," "Out-of-distribution Prompting" "Chain-Of-Thought Prompting," "N-shot Prompting," "Large Language Models," "Mental Health Classification," and "Mental Health Reasoning," related to methods, types, and applications on PE for digital mental health (DMH). The articles were compiled in a spreadsheet and then were filtered based on DMH type, PE type, PE method, PE task, LLMs used, and input data. To our knowledge, this is the first such review of PE-based methods for DMH. We summarize the overall review in Table 1 and discuss types of PE in Section 2. PE-based methods for DMH and applications are presented in Sections 3 and 4, respectively. Limitations, challenges, and future directions are described in Section 5.

# 2 Types of prompt engineering

## 2.1 N-shot prompting

N-shot prompting is an NLP technique for guiding LLMs to perform specific tasks with "*N*" examples in the prompt to understand the task. It enables in-context learning of LLMs for better performance with minimal additional training (2). Based on the in-context ("*N*") examples provided to LLMs, n-shot prompting can be further separated into zero-shot prompting and few-shot prompting. Zero-shot prompting has shown some promising results in performing well-designed prompt-driven non-complex tasks such as information retrieval, language translations, and question answering, without corresponding task-specific examples where the model must rely on its pre-existing knowledge and the task description in the prompt (25). Recent studies such as (13) and Lamichhane (14) have shown the capability of zero-shot prompting in ChatGPT for depression and suicidal detection. Few-shot prompting, on the other hand, performs well in complex tasks such as custom text generation and domain-specific question answering, using in-context examples (typically between two and five) along with task-specific prompts to steer an LLM for better understanding the task and to produce more accurate and contextually appropriate responses, compared to zero-shot prompting (26). Mental-RoBERTa (27) and Mental-FLAN-T5 (15) have been used to classify depression, stress, and suicidal thoughts using few-shot prompting.

## 2.2 Chain-of-thought (COT) prompting

COT prompting is an NLP technique to improve the reasoning capabilities of LLMs using structured prompts and immediate reasoning steps. In contrast with the application of LLMs to classification tasks using N-shot prompting, COT prompting helps the LLM to breakdown complex problems into manageable tasks and improves its ability to handle tasks using multi-step problem solving and explanation generation (28). Assessing the accuracy of LLM-generated explanations for mental health is critical. Kojima et al. (29) modified the vanilla prompt design using COT prompting to enhance the reasoning capability of GPT-3.5 and GPT-4 in metal health contexts. Englhardt et al. (30) suggested a novel approach based on multi-model time-series data to improve the reasoning abilities of LLMs for detecting depression and anxiety. A few studies have shown the explainability of LLMs in the context of mental health using end-user applications such as chatbots (31). Wang et al. (32) proposed a new COT framework to assess the mental status of users following multiple COT prompting reasoning steps in both zero-shot and few-shot settings. Chen et al. (33) introduced an enhanced version of COT prompting called Diagnosis of Thought prompting, a conceptual approach similar to COT prompting but focused more on

TABLE 1 Summary of the papers selected in this short review, classified into DHM type (D, depression; Anx, anxiety; ST, suicidal thoughts; CD, cognitive distortion; S, stress) PE types, PE task, data, and methods.

| Papers | DMH type | PE type | PE method | PE task | LLMs | Data | Results |
|---|---|---|---|---|---|---|---|
| Czejdo et al. (8) | D, Anx | N-shot COT | ICL | In-domain | GPT-3 Davinci | Q&A Summarization | Davinci's capability in n-shot Q&A |
| Tlachac et al. (9) | D, Anx | N-shot | ICL | In-domain | GPT-3 | Scripted audio (SA) Unscripted audio (USA) | D F1 (SA)—0.746 D F1 (USA)—0.691 Anx F1 (SA)—0.667 Anx F1 (USA)—0.63 |
| Ji (10) | ST | N-shot | IPT | In-domain | BERT MBERT | Reddit posts Weibo posts | F1 (BERT)—0.571 F1 (MBERT)—0.61 |
| Qi et al. (11) | ST, CD | N-shot | ICL PT | In-domain | GLM GPT-3.5 GPT-4 | Weibo posts Zoufan blogs | ST F1 (GLM)—0.722 ST F1 (GPT-4)—0.75 CD F1 (GLM)—0.17 CD F1 (GPT-4)—0.32 |
| Yang et al. (12) | D, S, ST | N-shot COT | ICL PT | In-domain | ChatGPT GPT-3 LLaMA | Reddit posts CLPsych15 Dreaddit T-SID | D F1 (GPT-3)—0.831 S F1(ChatGPT)—0.85 ST F1 (LLaMA)—0.54 |
| Amin et al. (13) | D, ST | N-shot | ICL | In-domain | ChatGPT | Reddit posts Sentiment-140 | D Accuracy—0.855 ST Recall—0.912 |
| Lamichhane (14) | D, S, ST | Few-shot | ICL | In-domain | ChatGPT | Reddit posts Dreaddit | D F1—0.73 S F1—0.86 ST F1—0.37 |
| Xu et al. (15) | D, S, ST | N-shot | ICL | In-domain | GPT-4 FLAN-T5 LLaMA | DepSeverity SDCNL CSSRS | D F1 (GPT-4)—0.719 S F1 (FLAN-T5)—0.67 ST F1 (LLaMA)—0.72 |
| Guo et al. (16) | D | N-shot | ICL | In-domain | PTDD | DAIC-WOZ | Accuracy—0.69 F1—0.60 |
| Ghanadian et al. (17) | ST | N-shot | ICL | In-domain | ChatGPT | Reddit UMD | Accuracy—0.88 F1—0.73 |
| Yang et al. (18) | D | N-shot | ICL | In-domain | ChatGPT GPT-4 LLaMA MLLaMA | IMHI | F1 (ChatGPT)—0.71 F1 (GPT-4)—0.781 F1 (LLaMA)—0.615 F1 (MLLaMA)—0.83 |
| Qin et al. (19) | D | N-shot COT | ICL | Out-of-distribution | BERT ChatGPT GPT-3 | Weibo posts Twitter MDD | F1 (BERT)—0.587 F1 (ChatGPT)—0.79 F1 (GPT-3)—0.851 |
| Ramos et al. (20) | D | N-shot | ICL | In-domain | BERT GPT-3.5 | SetembroBR | F1 (BERT)—0.65 F1 (GPT-3.5)—0.66 |
| Zhang et al. (21) | D | N-shot | ICL | In-domain | BERT T5 FGPL | DAIC-WOZ | F1 (BERT)—0.7407 F1 (T5)—0.75 F1 (FGPL)—0.7692 |
| Malhotra et al. (22) | D, Anx | N-shot | ICL | Out-of-distribution | BERT MBERT | Twitter posts | F1 (BERT)—0.866 F1 (MBERT)—0.888 |
| Agrawal (23) | D | COT | ICL | Out-of-distribution | GPT-4 LLaMA Gemini | DAIC-WOZ Reddit MHD | F1 (GPT-4)—0.74 F1 (LLaMA)—0.69 F1 (Gemini Pro)—0.66 |
| Chiu et al. (24) | S | N-shot | ICL | In-domain | GPT-3 GPT-3.5 GPT-4 | Therapy conversations HOPE | F1 (GPT-3)—0.496 F1 (GPT-3.5)—0.371 F1 (GPT-4)—0.577 |

understanding and validating the thought process behind the LLM's responses, to detect cognitive distortions. Although COT prompting improves the LLM's ability to handle complex tasks compared to N-shot prompting, the quality of prompts can limit the effectiveness.

## 3 Methods of prompt engineering

### 3.1 In-context learning (ICL)

ICL is the simplest PE method to adapt the knowledge of GPT-3 to solve a new, semantically similar tasks without additional explicit training using in-context examples, also known as demonstrations, inspired by the knowledge transferability of the human brain to new tasks using few instructions (2). Liu et al. (34) showed the importance of dynamically retrieved demonstrations over random demonstrations for natural language generation (NLG) tasks. Hayati et al. (35) explored the few-shot capability of GPT-3 for depression detection using contextually similar demonstrations. Su et al. (36) further demonstrated the mental health reasoning capabilities of LLMs using a new ICL framework. Fu et al. (37) introduced a commonsense-based response generation method by enhancing the explainability of ChatGPT and T5 models in the context of mental health using domain-specific demonstrations. Recently (38), developed the *GoodTimes* app, a personalized conversational and storytelling tool for reminiscence therapy, using the ICL-based reasoning capabilities of SOTA NLP models. As shown in Table 1, ICL-based N-shot prompting shows significant results in depression, stress, and suicidal thought detection (14, 15). Even though multiple DMH studies have been conducted for contextually similar knowledge transfer using ICL-based techniques, adapting knowledge to contextually dissimilar tasks is yet to be achieved due to limitations such as the lack of relevant contextual cues, differences in dissimilar tasks structures, limited generalization of LLMs to transfer knowledge, and the complexity of creating effective prompts for contextually dissimilar tasks (39).

### 3.2 Prompt tuning (PT)

Considering the limitations of ICL, soft continuous prompts were proposed to enhance the in-context capability of GPT-3 to execute a new task by adapting a few parameters while keeping the majority of the LLM frozen (4). Blair et al. (40) introduced a few-shot PT-based domain transfer technique for named entity disambiguation in mental health news articles. Li et al. (41) suggested novel PT-based optimization methods and a reinforcement learning framework for GPT-4 which can be used for mental health related NLG tasks. Spathis et al. (42) used PT-based evaluation protocols such as zero-shot inference to work with temporal stress levels data. According to Table 1, PT-based N-shot prompting performs better than ICL-based N-shot and COT prompting in suicidal thoughts and cognitive distortion detection (11, 12). PT-based methods are still unstable for scaling LLMs even though such methods outperform ICL-based approaches due to optimization challenges, LLM complexity, and the absence of sufficient contextual information for LLM generalization (43).

### 3.3 Instruction prompt tuning (IPT)

Recently, IPT was introduced as a combination of ICT and PT to facilitate the knowledge transfer of contextually dissimilar tasks by concatenating soft continuous prompts of the source task with retrieved demonstrations of the target task (39). Singhal et al. (44) introduced the concept of an LLMs' transferability to unseen tasks in classification and NLG medical domains. Nguyen et al. (45) proposed a novel depression screening process based on out-of-domain knowledge transfer methods. Ji (10) introduced an NLP-based suicidal risk detection method based on the sentiment classification capability of LLMs. Gupta et al. (46) explored the LLMs' zero-shot performance on unseen dialogue-related NLG tasks and cross-task generalization in multiple dialogue settings. The same approach was further modified to enhance the cross-task generalization capability of GPT-3 on stress screening (47).

## 4 Applications

Downstream tasks and applications depending on the transferability of soft prompts use in-domain, out-of-distribution, and out-of-domain PE-based mechanisms (48). In-domain prompt transfer adapts an LLM to a specific task within the same domain while out-of-distribution focuses on selecting a different distribution of the same source corpus within the same domain settings (49). Out-of-domain, which is the latest research trend, facilitates transferring LLMs into contextually dissimilar NLP tasks in different domains. In this section, applications of PE in DMH including classification, generation, and question answering tasks are discussed.

### 4.1 Classification task

Anxiety detection, depression detection, and suicidality detection are the most cited application domains of the DMH classification task. Abd-Alrazaq et al. (31) were the first to present a scoping review for n-shot ICL-based prompt engineering techniques in DMH. EMU framework, compatible with passive modalities, was introduced to screen depression and anxiety and the corpus was made publicly available for research purposes (9). Amin et al. (13) analyzed the depression detection capability of ChatGPT using n-shot prompting. Yang et al. (12) explored mental health analysis across five tasks including depression classification and introduced a reliable annotation protocol using emotion-enhanced COT prompting. Mental-LLM was introduced as a SOTA LLM for depression and stress classification using GPT-3.5 and GPT-4 prompting (15). Qi et al. (11) showed suicidality detection in social media posts using zero-shot and few-shot ICL-based prompting. Guo et al. (16) invented a topic modelling framework for depression detection on low-resource data based on handcrafted n-shot prompting. Only a few studies focused on the quality of generated responses by ChatGPT for suicidality detection using n-shot prompting (17). Recently (20), investigated LLM prompting for DMH using large and noisy social media corpora.

### 4.2 Generation task

Considering the reasoning capabilities of LLMs, several generation-based tasks for DMH can be identified. Prompt-based generation is important to predict mental health conditions. Yang et al. (12) showed the sensitivity of LLMs for different input prompts such as *severe* and *very severe* in explainable mental health analysis while mitigating the consequences using few-shot prompting. LLaMA-2 was used as a text augmentation assistant in content generation for mental healthcare treatment planning (50). MentalLLaMA was invented to improve the interpretability of LLMs in DMH (18). Qin et al. (19) introduced a novel COT prompting approach for depression detection and reasoning using zero-shot and few-shot out-of-distribution, which are unseen samples within the same domain, settings. Recently, this was further enhanced using explainable LLM-based techniques to understand psychological state (22). Agrawal (23) improved the explainability and reasoning of the latest generative LLMs in depression analysis using a novel COT prompting framework. Inspired by the Generate, Annotate, and Learn (GAL) framework by (51), a novel suicidality detection framework was introduced to generate synthetic data using LLMs to improve explainability (52). In comparison with classification-based tasks, most of the generation-based tasks use COT prompting as the PE type.

### 4.3 Question answering task

Only a few recent studies have demonstrated question-answering in psychological consultation services and online counselling for mental health professionals. Frameworks such as Psy-LLM, pretrained with LLMs and prompt-tuned with question-answering from psychologists, provide peer support and mental health advice in psychological consultation (53). Liu et al. (54) presented ChatCounselor, an enhanced LLM-based chatbot fine-tuned with domain-specific prompts and demonstrations to reinforce high-quality reasoning and question-answering in DMH. Recently (24), introduced BOLT, an ICL-based framework, to characterize the conversational behavior of clients and therapists.

# 5 Discussion

In this paper, we conducted a short review of how the latest prompt engineering methods in the context of digital mental health are being applied. We discussed three major application tasks to support DMH selecting two major types of PE, n-shot prompting and COT prompting, on ICL, PT, and IPT prompting methods introduced within last five years. In this section, we discuss the challenges, limitations, and future directions in PE for DMH.

There are a few challenges and limitations of PE for DMH. The primary challenge is the scarcity of the data needed to design relevant, accurate, and effective prompts for specific tasks in low-resource and cross-domain settings resulting in low performance during N-shot prompting-based classification and COT prompting-based generation tasks. A few publicly available datasets exist for some PT-based DMH tasks such as bipolar disorder detection, which require specific prompt designs. Even though a few recent studies attempted to mitigate the issue of data scarcity in PT-based tasks using low-resource and cross-domain settings, significant performance is yet to be achieved (55). Designing multiple prompts to improve the performance of N-shot prompting and selecting the most appropriate demonstrations for PT-based and IPT-based knowledge transferring to DMH applications can lead to higher computational requirements resulting scalability issues in LLMs. Although multiple studies recommend soft prompts over handcrafted prompts, it was found that the performance of LLMs tend to overfit due to the nature of bias in soft prompts (56). On the other hand, designing handcrafted prompts requires vast domain knowledge, clinical expertise, and terminology, resulting in uncertainty about better prompt designs for different N-shot prompting-based DMH tasks. In some cases, the performance of LLMs is over-estimated due to in-context information leakage and biased prompts (56). Another challenge is to select the most appropriate demonstrations for cross-model and cross-task transfer using different source and target prompts, to achieve LLM generalization for unseen data in N-shot ICL and COT prompting. Adapting the knowledge of a LLM for depression classification into a different task such as IPT-based depression reasoning is challenging due to the selection of effective demonstrations (57).

Prompt variability and framing plays an important role in maintaining the accuracy and reliability of LLMs in PT-based classification and generation-based tasks (58). An LLMs' probability of generating different predictions for a specific task is high due to the prompt framing effect. A few vulnerability attacks such as prompt leaking and goal hijacking expose confidential details to public scrutiny, by twisting the original task of a prompt, and this must be carefully prevented in DMH COT prompting-based reasoning tasks (59). Preventing adversarial attacks, manipulating LLMs to generate erroneous results using crafted prompts, is also a challenging task even though few attempts have been made to mitigate those using PT-based methods (60). Improving LLMs' interpretability and self-consistency in generation and reasoning tasks in DMH is also identified as a formidable challenge due to its complexity (61).

Using PT-based and ICL-based methods to work with mental health data brings several ethical considerations that need to be carefully addressed. An ethical-legal guidance and clinical validation framework is important to reduce the uncertainty in algorithmic bias, DMH data misuse and to improve LLM transparency and explainability (62). Data anonymization methods and carefully designed prompts should be used to improve the contextual understanding of LLMs mitigating privacy, confidentiality, uncertainty, and accountability issues in ICL-based reasoning. Model reliability should be validated when applying PT-based techniques to improve frozen LLM in-domain knowledge transferability for DMH tasks. Psychological impact and professional autonomy of clinical practitioners, on the other hand, should be carefully considered to assess the quality of prompt designs and in-context examples used for IPT-based out-of-domain DMH tasks.

Prompt automation and intelligence, automating downstream tasks using prompt-driven conversational agents, is a potential direction to enhance the efficiency and accuracy of DMH tasks by processing data more accurately (63). Multimodal COT prompting is an emerging trend to use COT prompting methods for processing multiple forms of mental health data such as text and images to further improve the reasoning capabilities (64). Recently, domain generalization for few-shot settings has been achieved to adapt learned prompts into unseen domains (65). Future research, such as pairing source task prompt embeddings with the in-context demonstrations of another different task and domain shifts with multiple soft prompts, is needed to achieve satisfactory performance in out-of-domain IPT-based task transfer.

## Author contributions

YP: Formal Analysis, Investigation, Validation, Writing – original draft, Writing – review & editing. AS: Formal Analysis, Investigation, Validation, Writing – original draft, Writing – review & editing. ZL: Supervision, Writing – review & editing. IP: Supervision, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

1. Qin Y, Lin Y, Yi J, Zhang J, Han X, Zhang Z, et al. Knowledge Inheritance for Pre-Trained Language Models. (2021). arXiv 2105.13880.

2. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. (2020) 33:1877–901. doi: 10.5555/3495724.3495883

3. Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, et al. GPT understands, too. *AI Open*. arXiv:2103.10385 (2023). doi: 10.1016/j.aiopen.2023.08.012

4. Lester B, Al-Rfou R, Constant N. The Power of Scale for Parameter-Efficient Prompt Tuning. (2021). arXiv 2104.08691.

5. Egan K. Digital technology, health and well-being and the COVID-19 pandemic: it's time to call forward informal carers from the back of the queue. *Semin Oncol Nurs*. (2020) 36(6):151088. WB Saunders. doi: 10.1016/j.soncn.2020.151088

6. William D, Suhartono D. Text-based depression detection on social media posts: a systematic literature review. *Procedia Comput Sci*. (2021) 179:582–9. doi: 10.1016/j.procs.2021.01.043

7. Liu H, Zhang W, Xie J, Kim B, Zhang Z, Chai Y. Few-Shot Learning for Chronic Disease Management: Leveraging Large Language Models and Multi-Prompt Engineering with Medical Knowledge Injection. (2024). arXiv 2401.12988.

8. Czejdo C, Bhattacharya S. Towards language models for AI mental health assistant design. In: *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*. Las Vegas, NV: IEEE. (2021). pp. 1217–22. doi: 10.1109/CSCI54926.2021.00252

9. Tlachac ML, Toto E, Lovering J, Kayastha R, Taurich N, Rundensteiner E. Emu: early mental health uncovering framework and dataset. In: *20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Pasadena, CA: IEEE (2022). pp. 1311–8. doi: 10.1109/ICMLA52953.2021.00213

10. Ji S. Towards intention understanding in suicidal risk assessment with natural language processing. In: Goldberg Y, Kozareva Z, Zhang Y, editors. *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi: Association for Computational Linguistics (2022). pp. 4028–38. doi: 10.18653/v1/2022.findings-emnlp.297

11. Qi H, Zhao Q, Song C, Zhai W, Luo D, Liu S, et al. Evaluating the Efficacy of Supervised Learning vs Large Language Models for Identifying Cognitive Distortions and Suicidal Risks in Chinese Social Media. (2023). arXiv 2309.03564.

12. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards interpretable mental health analysis with large language models. In: Bouamor H, Pino J, Bali K, editors. *The 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics (2023) 6056–77. doi: 10.18653/v1/2023.emnlp-main.370

13. Amin MM, Cambria E, Schuller BW. Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intell Syst*. (2023) 38(2):15–23. doi: 10.1109/MIS.2023.3254179

14. Lamichhane B. Evaluation of Chatgpt for NLP-Based Mental Health Applications. (2023). arXiv 2303.15727.

15. Xu X, Yao B, Dong Y, Yu H, Hendler J, Dey AK. Leveraging Large Language Models for Mental Health Prediction Via Online Text Data. (2023). arXiv 2307.14385.

16. Guo Y, Liu J, Wang L, Qin W, Hao S, Hong R. A prompt-based topic-modeling method for depression detection on low-resource data. *IEEE Trans Comput Soc Syst*. (2023) 11(1):1430–9. doi: 10.1109/TCSS.2023.3260080

17. Ghanadian H, Nejadgholi I, Osman HA. ChatGPT for Suicide Risk Assessment on Social Media: Quantitative Evaluation of Model Performance, Potentials and Limitations. (2023). arXiv 2306.09390.

18. Yang K, Zhang T, Kuang Z, Xie Q, Ananiadou S. Mentalllama: Interpretable Mental Health Analysis on Social Media with Large Language Models. (2023). arXiv 2309.13567.

19. Qin W, Chen Z, Wang L, Lan Y, Ren W, Hong R. Read, Diagnose and Chat: Towards Explainable and Interactive LLMs-Augmented Depression Detection in Social Media. (2023). arXiv 2305.05138.

20. Ramos dos Santos W, Paraboni I. Prompt-Based Mental Health Screening from Social Media Text. (2024). arXiv arXiv-2401.

21. Zhang J, Guo Y. Multilevel depression status detection based on fine-grained prompt learning. *Pattern Recognit Lett*. (2024) 178:167–73. doi: 10.1016/j.patrec.2024.01.005

22. Malhotra A, Jindal R. XAI transformer based approach for interpreting depressed and suicidal user behavior on online social networks. *Cogn Syst Res*. (2024) 84:101186. doi: 10.1016/j.cogsys.2023.101186

23. Agrawal A. Illuminate: a novel approach for depression detection with explainable analysis and proactive therapy using prompt engineering. *Int J Psychiatry*. (2024). doi: 10.13140/RG.2.2.19773.03042

24. Chiu YY, Sharma A, Lin IW, Althoff T. A Computational Framework for Behavioral Assessment of LLM Therapists. (2024). arXiv 2401.00820.

25. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al. Stanford Alpaca: An Instruction-Following Llama Model. (2023).

26. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and Efficient Foundation Language Models. (2023). arXiv 2302.13971.

27. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. Mentalbert: Publicly Available Pretrained Language Models for Mental Healthcare. (2021). arXiv 2110.15621.

28. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst*. (2022) 35:24824–37. arXiv 2201.11903. doi: 10.48550/arXiv.2201.11903

29. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. *Adv Neural Inf Process Syst*. (2022) 35:22199–213. arXiv abs/2205.11916. doi: 10.48550/arXiv.2205.11916

30. Englhardt Z, Ma C, Morris ME, Xu X, Chang CC, Qin L, et al. From Classification to Clinical Insights: Towards Analyzing and Reasoning about Mobile and Behavioral Health Data with Large Language Models. (2023). arXiv 2311.13063.

31. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. *J Med Internet Res*. (2021) 23(1):e17828. doi: 10.2196/17828

32. Wang H, Wang R, Mi F, Wang Z, Xu R, Wong KF. Chain-of-Thought Prompting for Responding to in-Depth Dialogue Questions with LLM. (2023). arXiv 2305.11792.

33. Chen Z, Lu Y, Wang WY. Empowering Psychotherapy with Large Language Models: Cognitive Distortion Detection Through Diagnosis of Thought Prompting. (2023). arXiv 2310.07146.

34. Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What Makes Good in-Context Examples for GPT-3?. (2021). arXiv 2101.06804.

35. Hayati MFM, Ali MAM, Rosli ANM. Depression detection on Malay dialects using GPT-3. In: *2022 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. Kuala Lumpur: IEEE. (2022). pp. 360–4. doi: 10.1109/IECBES54088.2022.10079554

36. Su H, Kasai J, Wu CH, Shi W, Wang T, Xin J, et al. Selective Annotation Makes Language Models Better Few-Shot Learners. (2022). arXiv 2209.01975.

37. Fu Y, Inoue K, Chu C, Kawahara T. Reasoning Before Responding: Integrating Commonsense-Based Causality Explanation for Empathetic Response Generation. (2023). arXiv 2308.00085.

38. Wang X, Li J, Liang T, Hasan WU, Zaman KT, Du Y, et al. Promoting personalized reminiscence among cognitively intact older adults through an AI-driven interactive multimodal photo album: development and usability study. *JMIR Aging*. (2024) 7(1):e49415. doi: 10.2196/49415

39. Sun S, Liu Y, Iter D, Zhu C, Iyyer M. How Does in-Context Learning Help Prompt Tuning? (2023). arXiv 2302.11521.

40. Blair P, Bar K. Improving few-shot domain transfer for named entity disambiguation with pattern exploitation. In: Goldberg Y, Kozareva Z, Zhang Y, editors. *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi: Association for Computational Linguistics (2022). pp. 6797–810. doi: 10.18653/v1/2022.findings-emnlp.506

41. Li C, Liu X, Wang Y, Li D, Lan Y, Shen C. Dialogue for Prompting: A Policy-Gradient-Based Discrete Prompt Optimization for Few-Shot Learning. (2023). arXiv 2308.07272.

42. Spathis D, Kawsar F. The First Step is The Hardest: Pitfalls of Representing and Tokenizing Temporal Data for Large Language Models. (2023). arXiv 2309.06236.

43. Ding N, Qin Y, Yang G, Wei F, Yang Z, Su Y, et al. Delta Tuning: A Comprehensive Study of Parameter Efficient Methods for Pre-Trained Language Models. (2022). arXiv 2203.06904.

44. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large Language Models Encode Clinical Knowledge. (2022). arXiv 2212.13138.

45. Nguyen T, Yates A, Zirikly A, Desmet B, Cohan A. Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires. (2022). arXiv 2204.10432.

46. Gupta P, Jiao C, Yeh YT, Mehri S, Eskenazi M, Bigham JP. Improving Zero and Few-Shot Generalization in Dialogue Through Instruction Tuning. (2022). arXiv 2205.12673.

47. Mishra S, Nouri E. Help Me Think: A Simple Prompting Strategy for Non-Experts to Create Customized Content with Models. (2023). arXiv 2208.08232.

48. Su Y, Wang X, Qin Y, Chan CM, Lin Y, Wang H, et al. On Transferability of Prompt Tuning for Natural Language Processing. (2021). arXiv 2111.06719.

49. Vu T, Wang T, Munkhdalai T, Sordoni A, Trischler A, Mattarella-Micke A, et al. Exploring and Predicting Transferability Across NLP Tasks. (2020). arXiv 2005.00770.

50. Bhaumik R, Srivastava V, Jalali A, Ghosh S, Chandrasekaran R. Mindwatch: a smart cloud-based ai solution for suicide ideation detection leveraging large language models. *medRxiv*. (2023). medRxiv 2023.09.25.23296062. doi: 10.1101/2023.09.25.23296062

51. He X, Nassar I, Kiros J, Haffari G, Norouzi M. Generate, annotate, and learn: NLP with synthetic text. *Trans Assoc Comput Linguist*. (2022) 10:826–42. doi: 10.1162/tacl_a_00492

52. Ghanadian H, Nejadgholi I, Al Osman H. Socially aware synthetic data generation for suicidal ideation detection using large language models. *IEEE Access*. (2024) 12:14350–63. doi: 10.1109/ACCESS.2024.3358206

53. Lai T, Shi Y, Du Z, Wu J, Fu K, Dou Y. Psy-LLM: Scaling Up Global Mental Health Psychological Services with AI-Based Large Language Models. (2023). arXiv 2307.11991.

54. Liu JM, Li D, Cao H, Ren T, Liao Z, Wu J. Chatcounselor: A Large Language Models for Mental Health Support. (2023). arXiv 2309.15461.

55. Wang C, Yang Y, Gao C, Peng Y, Zhang H, Lyu MR. No more fine-tuning? An experimental evaluation of prompt tuning in code intelligence. In: *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. Singapore and New York, NY: Association for Computing Machinery (2022). pp. 382–94. doi: 10.1145/3540250.3549113

56. Cao B, Lin H, Han X, Sun L, Yan L, Liao M, et al. Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases. (2021). arXiv 2106.09231.

57. Xu X, Liu Y, Pasupat P, Kazemi M. In-Context Learning with Retrieved Demonstrations for Language Models: A Survey. (2024). arXiv 2401.11624.

58. Wang W, Haddow B, Birch A, Peng W. Assessing the Reliability of Large Language Model Knowledge. (2023). arXiv 2310.09820.

59. Perez F, Ribeiro I. Ignore Previous Prompt: Attack Techniques for Language Models. (2022). arXiv 2211.09527.

60. Zou A, Wang Z, Kolter JZ, Fredrikson M. Universal and Transferable Adversarial Attacks on Aligned Language Models, 2023. (2023). arXiv 2307.15043.

61. Li X, Xiong H, Li X, Wu X, Zhang X, Liu J, et al. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowl Inf Syst*. (2022) 64(12):3197–234. arXiv 2103.10689. doi: 10.1007/s10115-022-01756-8

62. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. *Front Digit Health*. (2021) 3:697072. doi: 10.3389/fdgth.2021.697072

63. Martin EA, D'Souza AG, Lee S, Doktorchik C, Eastwood CA, Quan H. Hypertension identification using inpatient clinical notes from electronic medical records: an explainable, data-driven algorithm study. *Can Med Assoc Open Access J*. (2023) 11(1):E131–9. doi: 10.9778/cmajo.20210170

64. Zhang D, Yang J, Lyu H, Jin Z, Yao Y, Chen M, et al. Cocot: Contrastive Chain-of-Thought Prompting for Large Multimodal Models with Multiple Image Inputs. (2024). arXiv 2401.02582.

65. Zhao C, Wang Y, Jiang X, Shen Y, Song K, Li D, et al. Learning domain invariant prompt for vision-language models. *IEEE Trans Image Process*. (2024) 33:1348–60. doi: 10.1109/TIP.2024.3362062