Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cyberbetics

Neuro-Computing Methods for Major Depressive Disorder Detection and Psychotherapy Aid

Disertation Thesis

Cheng Kang

Ph.D. programme: Bioengineering
Supervisor: Doc. Ing. Daniel Novak, Ph.D.

Prague, January 2024

Thesis Supervisor:
Doc. Ing. Daniel Novak, Ph.D.
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague
Technická 2
160 00 Prague 6
Czech Republic

Declaration

I hereby declare I have written this doctoral thesis independently and quoted all the sources of information used in accordance with methodological instructions on ethical principles for writing an academic thesis. Moreover, I state that this thesis has neither been submitted nor accepted for any other degree. The results presented in this dissertation have been published in [1]–[6] during my Ph.D. research in cooperation with my dissertation supervisor Daniel Novak. In my Ph.D. study, I collaborated with several researchers on multiple projects. I publish articles with Yong Hu [1], [6]–[9], Yuezhi Li [1], [7], [8], Huiyu Zhou [9], Yudong Zhang [1]–[4], Xujing Yao [4], [5], Jindrich Prokop [9], Xiang Yu [3].

In Prague, January 2024

....
Cheng Kang

Abstract

The detection of Major Depressive Disorder (MDD) has benefited from advanced neurocomputing methods and traditional machine learning techniques. In addition, new technical tools have been trying to relieve patients' suffering. In general, studies about the detecting rate of depression are mostly too low to be transferred to clinical applications, and techniques about psychological therapies or assistance are heavily relying on specific places and times. In this thesis, I describe the results of three projects that address challenges to making the depression detection rate more stable with a higher accuracy rate and making an available psychotherapy chatbot without pretraining on huge language datasets but with a stronger performance. In the second chapter, I and my co-authors developed a Brain-Computer Interface (BCI) system for processing electroencephalogram (EEG) signals and constructing the dynamic functional brain networks between depressive patients and healthy controls. Meanwhile, two residual neural networks based on selected EEG channels and frequencies were used to detect depression from the health and to evaluate the depressive severity with the score of Structured Clinical Interview for DSM-IV Axis I Disorders, Clinician Version (SCID-CV). In the third chapter, I and my co-authors proposed a novel Fuzzy Window with the Gaussian Processed Labels (FW-GPL) method for ordinal scoring tasks. With the use of window process, this model has the advantage to process ordinal data, such as, medical images and EEGs of patients with different depressive severity. In the forth chapter, to develop advanced training or fine-tuning methods based on neuroscience knowledge, I and my co-authors studied the brain functional dynamics during Working Memory (WM), and we found maintenance, inhibition and disinhibition should work together to process the information in our brain. Depends on these findings in chapter three, we proposed a neuroscience-inspired architecture model, shunting inhibition in chapter four, and the results of this new architecture on fine-tuning downstream language tasks prove the effectiveness of gating Multilayer Perceptions (MLPs) and inhibition mechanisms. In the fifth project, I and my co-authors developed a psychotherapy chatbot fine-tuned on Large Language Models (LLMs) processed AlexanderStreet therapy and counseling data, and it provided more professional and common used psychotherapy knowledge. Aside from contributing scientific conclusions about each system, these methods will also serve as a practical framework for future efforts to address challenges to depression detection and psychotherapy aid.

Keywords: Depression detection, depressive severity scoring, brain computer interface, ordinal scoring tasks, parameter efficient fine tuning, large language models, psychotherapy chatbot.

Abstrakt

Dizertační práce se zaměřuje na detekci deprese s využitím pokročilých neurovědeckých výpočetních metod a tradičních technik strojového učení. Většina studií zaměřená na detekci deprese dosahuje příliš nízké přesnosti pro nasazení v klinické praxi a metody v oblasti psychologické terapie nebo asistence jsou značně závislé na specifických místech a časech. V této disertační práci prezentuji výsledky experimentů, které se zaměřují na zvýšení přesnosti detekce deprese a na vývoj přístupného psychoterapeutického chatbota, který nevyžaduje předchozí trénink na rozsáhlých jazykových datech. Ve druhé kapitole jsem vyvinul systém pro analýzu signálů elektroencefalogramu (EEG). Byly použity dvě reziduální neuronové sítě, jejich vstupem byly specifické EEG kanály a frekvenční spektrum signálu. Hodnocení závažnosti deprese bylo provedeno pomocí metodiky Structured Clinical Interview for DSM-IV Axis I Disorders, Clinician Version (SCID-CV). Ve třetí kapitole jsem navrhl inovativní metodu Fuzzy Window with Gaussian Processed Labels (FW-GPL) pro ordinální skórovací úlohy. Díky procesu oken tento model efektivně zpracovává např. lékařské snímky a EEG signály. Ve čtvrté kapitole jsem se zaměřil na zkoumání funkční dynamiky mozku během operací v pracovní paměti (WM) a zjistil jsem, že procesy udržování, inhibice a disinhibice vyvíjejí součinnost při zpracování informací v mozku. Ve čtvrté kapitole jsem navrhl neurovědami inspirovaný architektonický model založený na inhibici, jehož výsledky v jemném doladění jazykových úloh dokazují zvýšení přesnosti odpovědí jazykových modelů. V páté kapitole jsem vyvinul psychoterapeutického chatbota, který byl optimalizován na datech z AlexanderStreet knihovny. Předložené metody budou sloužit jako praktický rámec pro budoucí snahy řešit výzvy v detekci deprese a v následné podpoře v psychoterapii.

Klíčová Slova: Detekce deprese, skórování závažnosti deprese, mozkové počítačové rozhraní, ordinální skórovací úlohy, parametricky efektivní jemné ladění, velké jazykové modely, psychoterapeutický chatbot.

Acknowledgements

I would like to express my heartfelt gratitude to all those who have contributed to the completion of my PhD thesis. This journey has been challenging, yet immensely rewarding, and I couldn't have reached this milestone without the support, encouragement, and assistance of numerous individuals.

- First and foremost, I am deeply thankful to my advisor, Daniel Novak, whose unwavering guidance, wisdom, and mentorship have been invaluable throughout my doctoral studies. Your dedication to my academic and personal growth has been instrumental in shaping the researcher I have become.
- Tomáš Sieger, Eduard Bakštein, Jiří Anýž, Jakub Schneider and fellow PhD student, for joining discussion and helping me with some tedious task.
- Jindřich Prokop, Ihor Varha and Václav Burda, Xujing Yao, Lei Tong, Xiang Yu, these PhD students, for being always available to listen and share their opinion, not minding the time I have stolen from them.
- I would like to acknowledge the faculties and staffs in University of Hong Kong, University of Leicester and Shenzhen University, whose commitment to academic excellence provided a nurturing environment for my research endeavors. The resources, facilities, and intellectual stimulation provided by the university were indispensable. Yuezhi Li, Yong Hu, Huiyu Zhou and Yudong Zhang professors of computer-science and neuroscience, for answering a lot of rather naive questions and letting me work in such an interesting field by providing computer science knowledge on medical domain and by steering the team towards relevant problems.
- Lastly, I dedicate this work to Yuqing Chen, Patrik Jankuv and Fabián Bodnar, for helping me gather necessary research data.
- I am grateful to my family, my father Linchao Kang, my mother Xiuge Lei and my girlfriend Qingyun Yang, for their unwavering love and encouragement. Your belief in me sustained my determination during the highs and lows of this academic pursuit.

  Thank you all for being an integral part of this academic journey. Your contributions, whether big or small, have left an indelible mark on my life.

Contents

List of Tables

List of Figures

+

Chapter 1

Introduction

The introduction first describes goals of this thesis, and depression as well as its common symptoms. Then traditional diagnosis and therapy ways are presented. In the second, more stable depression detection techniques are presented. Following the above, the introduction also describes one challenge when detecting depressive severity, which could seriously influence the detection and scoring results. To provide convenient and accessible tools which contains professional psychotherapy knowledge at any time to remind users, we firstly present one advanced Parameter-Efficient Tuning Method (PEFT) which was inspired by inhibition mechanism in our brain. Secondly, a short introduction to Nature Language Processing (NLP) on psychotherapy chatbot using LLMs is given with a more detailed summary in the following chapters. Lastly, a brief description of the goals and the thesis outline is given.

1.1 Goals of the Thesis

As shown in Figure 1.1, this thesis aims to develop a system that can detect depressive severity using brain computer interface and provide psychotherapy assistant using language models. Considering clinical knowledge of depression detection, and applying widely-used advanced Artificial Neural Networks (ANNs), this system can achieve an acceptable result in detecting depression and scoring depressive severity (in Chapter 2). After the detection and severity scoring procedure, more stable and advanced algorithms, such as fuzzy windows and Gaussian processed labels are proposed to improve the performance of detecting ordinal samples (in Chapter 3). Next, inspired by the inhibition and disinhibition brain networks (in Chapter 4) a better fine-tuning method that tunes pretrained language models - inhibited gate MLPs - was designed to improve the performance on specific downstream tasks (in Chapter 5). Finally, the psychotherapy data which was revised and augmented by GPT-4 can teach other LLMs to generate an efficient and reliable response (in Chapter 6). The primary objectives of this thesis can be categorized into:

- to analyze the abnormal brain connections using brain computer interface.
- to detect depressive severity using brain computer interface and automatic diagnosis system.
- to develop a better ordinal regression model, and then, to improve the detecting rate of depressive severity, aw well as other ordinal datasets.
- to develop a better fine-tuning method to tune the pre-trained large language models on professional knowledge.
- to develop a psychotherapy chatbot that can provide professional assistant to clients using large language models.

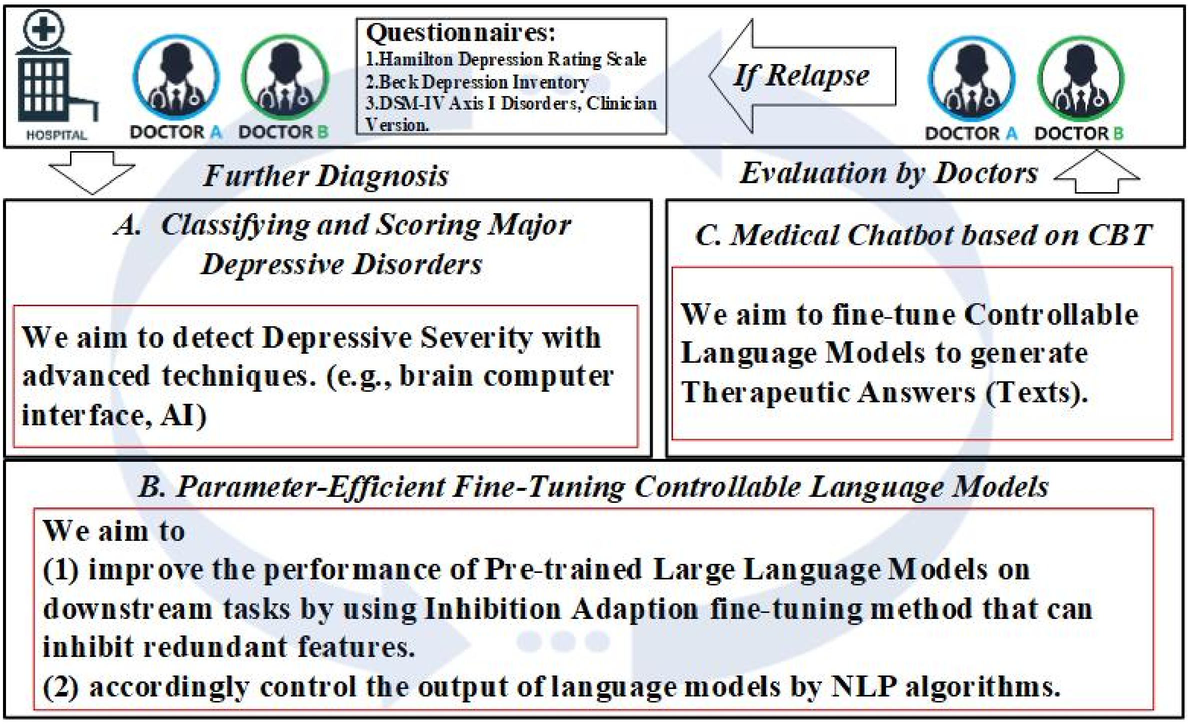


Figure 1.1: The whole framework of depression detection and assistance in this thesis.

1.2 Background

Depression also called MDD is a common illness worldwide, as reported by World Health Organization (WHO), with an estimated 3.8% of the population affected, including 5.0%

among adults and 5.7% among adults older than 60 years, and there are approximately 280 million people in the world have depression [10]. Depression is different from usual mood fluctuations and short-lived emotional responses to challenges in everyday life. Especially when recurrent and with moderate or severe intensity, depression may become a serious health condition. It can cause the affected person to suffer greatly and function poorly at work, at school and in the family. At its worst, depression can lead to suicide. Over 700 000 people die due to suicide every year. Suicide is the fourth leading cause of death in 15-29-year-olds. Although there are known, effective treatments for mental disorders, more than 75% of people in low- and middle-income countries receive no treatment [11]. Barriers to effective care include a lack of resources, lack of trained healthcare providers and social stigma associated with mental disorders. In countries of all income levels, people who experience depression are often not correctly diagnosed, and others who do not have the disorder are too often misdiagnosed and prescribed antidepressants. MDD is a mental illness which is often accompanied by a high risk of suicidal thoughts [12]. Depressed individuals are often misdiagnosed by physicians, which leads to a range of problems, including self-medication, substance abuse, inappropriate treatment, social isolation, and impaired performance in education or at work [13], [14]. Cognitive behavioural therapy is the best way to treat mild depression, and for severe depression, currently, the combination of psychotherapy and antidepressant drugs is the most effective treatment [15]–[17]. Improper treatments would lead to future relapse and prolonged discontinuation symptoms [18].

1.2.1 Detecting Depression

Challenges of Detecting Depression and Scoring Depressive Severity

Depression is widely categorized as non-depressed, mild, moderate, and severe, according to the severity of the depressive symptoms [19]. However, a descriptive study has shown that the rate of misdiagnosis of MDD is as high as 65.9% [14]. This means that the primary accuracy rate is less than 35% [14]. Failure to correctly diagnose MDD is caused by inadequate training of clinicians, as well as reasons that sufferers are not given appropriate appointments, medical examinations and proper treatments at the early stage [14], [20]. Existing tools for diagnosing MDD tend not to be used by clinical psychologists and physicians because these complex approaches have three main challenges:

(1) they are time-consuming and need to be administrated by well-trained engineers or by professional clinicians [21], [22]; **(2)** they cannot classify depressive severity; **(3)** there is no visualization result provided, for example, brain topological maps. The techniques used for depressive disorder detection can be divided into three rough categories: (**1**) ques-

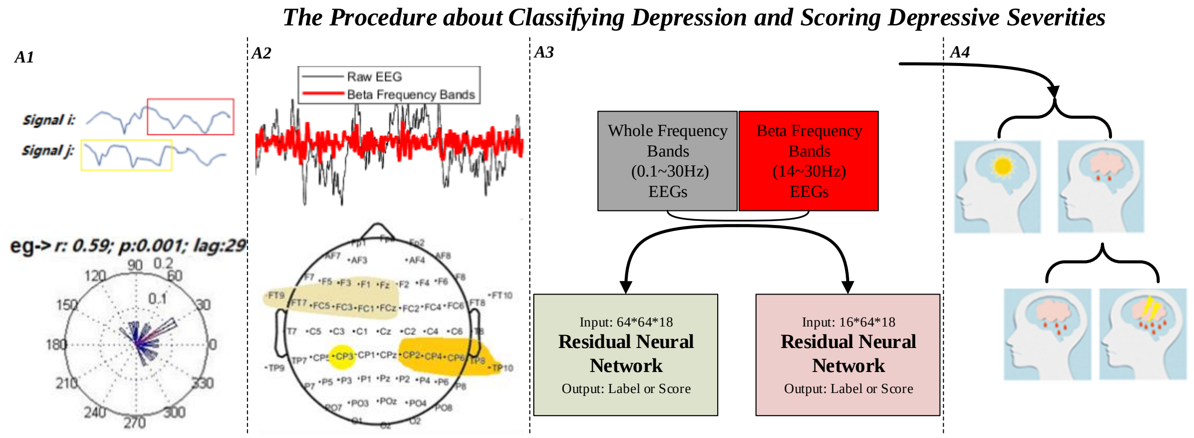


Figure 1.2: The framework of depressive severity scoring system using a BCI system. The entire procedure about classifying depression and scoring depressive severity (**A1** → **A2** → **A3** → **A4**).

tionnaires, (**2**) clinical sensors and (**3**) ubiquitous sensors. There are three most popular **questionnaires**: the SCID-CV [23], the Hamilton Depression Rating Scale (HAMD) [24], and the Beck Depression Inventory (BDI) [25], and all these three have strong histories of use in the psychological sciences. However, the obvious shortage is that this detection method heavily relies on the knowledge and experience of psychologists. Thus, this will inevitably bring much manual interference into depression detection. There are three most popular **clinical diagnosis and analysis methods**: electrodes, Magnetic Resonance Imaging (fMRI) [26], and Functional near-infrared spectroscopy (fNIRS) [27], [28]. The electrode techniques can provide several main physiological parameters: electrocardiogram (ECG) [29] and EEG [30], [31]). But the obvious shortage is the time-consuming process. There are a huge amount of **ubiquitous sensors** that are used outside of hospitals, allowing for mood disorder detection in any location. These sensors include fashion devices most people are familiar with, such as cameras [32], [33], smartphones [34], GPS [35], and WiFi [36], as well as sensors that work in the background of devices, such as accelerometers [37] and phone metadata. But these techniques require more clinical experiments to verify their reliability and dependability.

Challenges of Ordinal Scoring on Depressive Severity

Ordinal scoring, typically known as ordinal classification, is a supervised learning problem aiming to predict a discrete set of ordinal labels. The main difference from the classification task is that the categories are related in a natural or implied order. Ordinal classification can be viewed as a special case of metric regression, where the regression targets are discrete and finite. The depressive severity based on the HAMD and SCID-

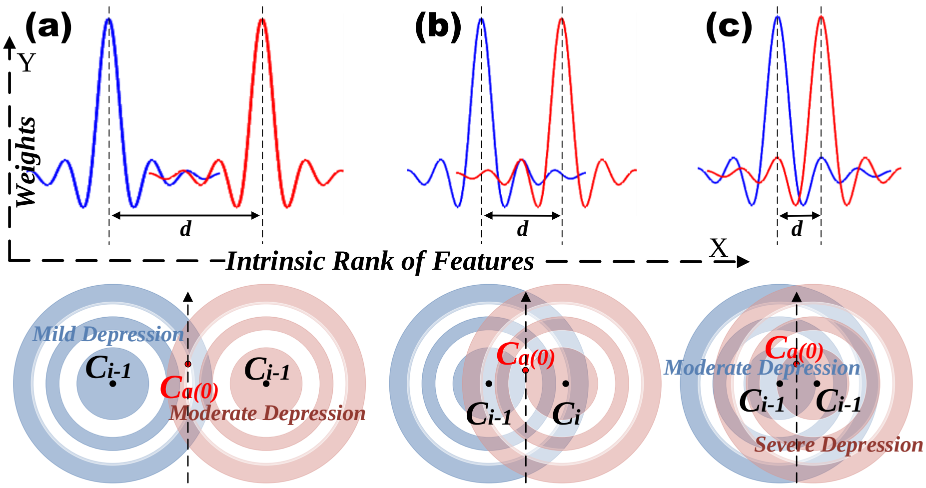


Figure 1.3: The challenge of ordinal image classification (or scoring). The X-axis denotes the intrinsic rank of features, and the Y-axis denotes the weights of models.

CV was classified, and the 17-item HAMD cut-off points were also defined as follows: $> 24 = severe$; $17 \sim 23 = moderate$; $8 \sim 16 = mild$; and $none(non-depressed) = 0 \sim 7$ [38]. The differences in features between adjacent labels are not always equal to each other. The difference in depressive symptoms between "Mild Depression" and "Moderate Depression" being more obvious than that between "Moderate Depression" and "Sever Depression" is one example. However, if the ordinal relationship of labels is ignored, the ordinal regression problem will only become a simple multi-class classification issue. We assume that the features in ordinal labels have an "intrinsic rank," and the corresponding ordinal category will show a specific concentration in terms of the "intrinsic rank." C_i and C_{i-1} are, respectively, the centers of their corresponding neighboring ordinal classes. (**a**) If the distance d between two centers is remote, the "intrinsic rank" is slack. (**b**) If the distance d between two centers is approaching the boundary, the "intrinsic rank" is tight. (**c**) If the distance d between two centers is beyond the boundary, the "intrinsic rank" seems to become a whole part. Under this condition, the classification task would become extremely difficult. For example, when scoring depressive severity, the obstacle that two neighbour categories closely share the overlapped features should be tackled, and most time, we always could not further provide the sub-scores which can represent the detailed information. When learning with ordinal labels, a common problem is that the ambiguity between two neighboring categories usually has a negative effect on the training convergence. Therefore, the performance of the learned model tends to degrade in ordinal classes. This challenge has motivated us to develop a robust ordinal label classification approach to analyzing ordinal data.

Ordinal classification approaches or ordinal models [39] can be roughly divided into two aspects, Single Label Learning with Specific Loss (SLL-Loss) [40]–[43] and the Label

Distribution Based Learning (LDBL) [40], [44]–[51]. SLL-Loss methods typically rely on independently processing a single facial image. This ignores gradual changes in human faces, and thus, facial appearance is usually ambiguous as regards adjacent age classes. The LDBL methods tend to map ordinal ground-truth learning based on a Gaussian or Gaussian-like label distribution. But in such a long-tailed case, they also ignore the processing of ordinal neighbours or overlapping features.

1.2.2 Psychotherapy Using Large Language Models

Challenges of Fine-tuning Large Language Models

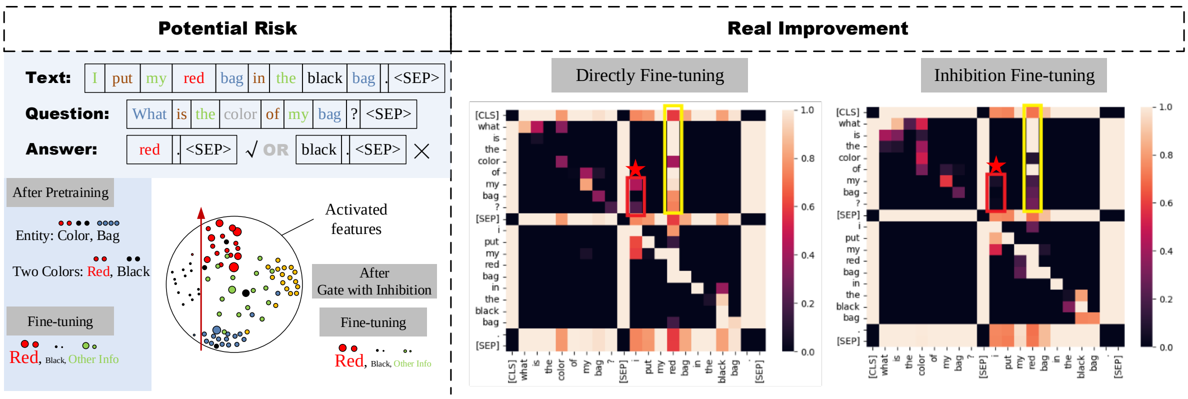


Figure 1.4: A practical example of InA and its use in the $BERT_{large}$ model, which has been fine-tuned under question-answering datasets.

Fine-tuning, the process of updating the parameters of pre-trained Language Models (LMs), has proven to be an effective approach for various downstream NLP tasks. However, classical fine-tuning methods suffer from the issue of redundant parameters in fully pre-trained models, which can lead to inefficiencies when adapting to new downstream tasks. To tackle this problem, prior studies have attempted to adapt only specific vectors or learn additional parameters while keeping most of the pre-trained parameters fixed. This allows for better operational efficiency by loading task-specific parameters associated with the pre-trained models before deployment. Low-Rank Adaption (LoRA) ([52] (has successfully achieved this goal and addressed the inference latency problem, which helps extend model depth or reduce the usable sequence length of models ([53]–[55]) to find a balance between efficiency and quality. The challenges in fine-tuning pre-trained LMs for Nature Language Understanding (NLU) downstream tasks lie in reducing the number of tuned weights and appropriately approximating the update of pre-trained weights derived from the LMs ([52], [53], [55], [56]). Properly selecting knowledge from pre-trained LMs is crucial to address these challenges. The question arises as to why we cannot directly inhibit ”redundant” knowledge during fine-tuning while retaining relevant information.

In the prior work of LoRA [52], authors only used the similarity matrix to compare the difference between LoRA fine-tuning and fully fine-tuning methods. There is no straight forward visualization result that can show us which part has been tuned by such methods. In addition, when using LoRA fine-tuning method on LMs, we found that although the low rank ”bottleneck” can compress information and reweight the pre-trained parameters, such compressed information always contains noise and task-irrelevant knowledge. As shown in Figure 1.4, we present an example: input = ['I put my red bag in the black bag. What is the colour of my bag ?'], target = ['red']. When the threshold is 0, InA will become to LoRA, as InA also uses low rank to compress the passing information. The target-irrelevant knowledge in this case includes pronouns (e.g., I, my, and what), nouns (e.g., bag), verbs (e.g., put), definite articles (e.g., the), and adjectives (e.g., black and colour). Both full fine-tuning (FT) and adaption FT methods still retain this target-irrelevant information, which can distract the model from focusing on the actual target knowledge. When the target is specified as ['red colour'], the relevant knowledge should be the adjective ”colour.” Figure 2 is a cross attention map, and it presents the “word connection” between the column and the row word lists. The “word connection” between “I” and “red” is reasonable, but the most important “word” should be “red”. To make attention layers pay more attention to most important “words”, that means making attention layers more concentrated, the noise words, such as “I” should be inhibited. Therefore, it is essential to eliminate such target-irrelevant information to ensure the model's output is more concentrated on the desired target. On the right-hand side of Figure 1.4, InA is introduced as a method to reduce the influence of the target-irrelevant knowledge, such as the pronoun ”I.”

Challenges of Developing Psychotherapy Chatbots Using Large Language Models

Large Language Models (LLMs) have demonstrated impressive generalization capabilities, such as in-context learning [57], chain-of-thoughts reasoning [58], and biomedical diagnosing [59]. Instruction-tuning of LLMs has enabled them to follow natural language instructions and perform real-world tasks [60]. Two main methods have been developed for instruction-tuning LLMs: **(1)** fine-tuning the model on a wide range of tasks using human-annotated prompts and feedback [61], and **(2)** supervised fine-tuning using public benchmarks and datasets augmented with manually or automatically generated instructions [62]. Reinforcement Learning on Human Feedback (RLHF) has proven to be an effective way to improve LLMs in various domains, such as medicine [63], knowledge graphs [64], and biomedical applications [65], but it comes with a high cost. Natural instructions

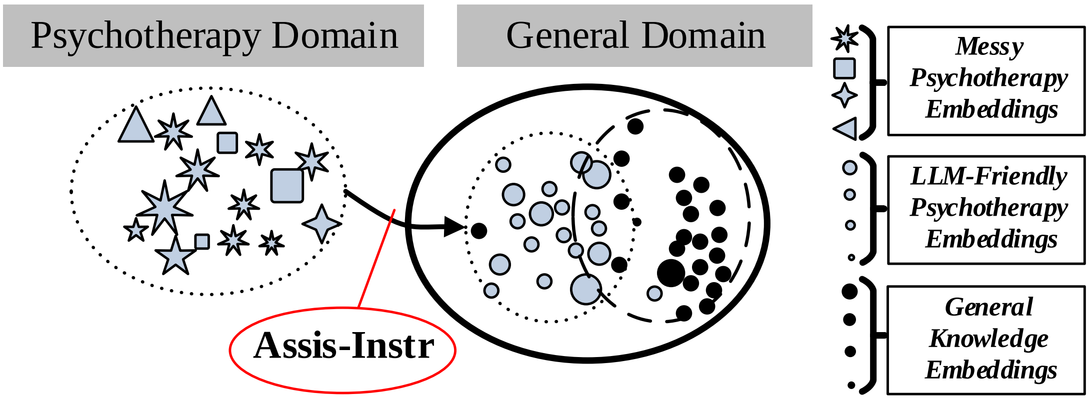


Figure 1.5: A semantic graph that describes how Assistant-Instruction can change the professional embedding to a common embedding. A successful model is expected to use the provided instructions (including task and domain definition examples) to output responses to professional evaluation instances.

[60], and even un-natural instructions [66], can provide knowledge in multiple domains, but LLMs pre-trained on vast corpora (e.g., Llama1 [67] and Llama2 [68] containing books, common crawled conversations, arxiv articles, GitHub, C4, and Wikipedia data) still require additional professional knowledge, especially from domain experts. Self-Instruct tuning [69], [70] and Guess-Instruction tuning methods have shown better performance in aligning LLMs with human intent by learning from instruction-following data generated by state-of-the-art instruction-tuned teacher LLMs (e.g., GPT-3, GPT-3.5, and even GPT-4). These lines of instruction-tuning research have proven effective in improving the zero and few-shot generalization abilities of LLMs.

The dataset we aim to generate consists of a collection of instructions $\{I_t\}$, where each instruction defines a specific domain t in natural language. Each domain t comprises $n_t \geq 1$ input-output instances $\{(X_{t,i}, Y_{t,i})\}_{i=1}^{n_t}$. We hypothesize that each domain t has its own distinct characteristics (as shown in the left panel of Figure 1.5). The objective is for a model M to generate the correct output based on the domain instruction and the corresponding input: $M(I_t, X_{t,i}) = Y_{t,i}$, for $i \in \{1, \ldots, n_t\}$. The instruction is formulated as "Provide suggestions or comments on addressing and alleviating the following topic," and the instance input is formatted as "addictive disorders." It is important to note that in some cases, there may not be strict boundaries between the instruction and instance input. For example, if the instruction is "Summarize the bellow description and explain the below concept on [***] domain. Add more common knowledge." and instance input is "Addiction and Spiritual Crisis.", the instruction domain may overlap with other domains. It may not always be possible to construct instructions (especially the output) that contain specific professional knowledge. Because multi-domain knowledge will make the training unstable, and the LLMs will generate the answer with some irrelevant knowledge. To promote diversity and individuality in the data format, we allow these instructions, instance inputs, and outputs to incorporate additional knowledge and assistant from other models (i.e., $Y = Y + Y'$, where Y' is revised by GPT-4 and then generated from GPT-4). In the right panel of Figure 1.5, we encounter the challenge of making the data LLM-friendly, wherein we use LLMs themselves to format instructions, instance inputs, and outputs.

1.3 Thesis Outline

In this thesis, I will explore the questions presented above, and hope to broaden our understanding of both the depressive severity detection aspects as well as the psychotherapy of depression using chatbots, a still relatively unexplored field. The structure of the thesis primarily relies on the content of these following five publications:

1. **Kang, C.***; Li, Y.*; Novak, D.; Zhang, Y.; Zhou, Q.; Hu, Y. (2020). Brain Networks of Maintenance, Inhibition and Disinhibition During Working Memory. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 7, pp. 1518-1527, July 2020.

2. **Kang, C.**; Novak, D.; Yao, X.; Xie, J.; Hu, Y. (2023). Classifying and Scoring Major Depressive Disorders by Residual Neural Networks on Specific Frequencies and Brain Regions. IEEE Trans Neural Syst Rehabil Eng. 2023;31:2964-2973.

3. **Kang, C.**; Yao, X.; Novak, D. (2023). Fuzzy Windows with Gaussian Processed Labels for Ordinal Image Scoring Tasks. Appl. Sci. 2023, 13, 4019.

4. **Kang, C.**; Prokop, J.; Tong, L.; Zhou, Z.; Hu, Y.; Novak, D. (2023). InA: Inhibition Adaption On Pre-trained Language Models. Submitted to Neural Networks (Minor Revision)

5. **Kang, C.**; Cheng, Y.; Zhang, Y.; Hu, Y.; Novak, D. (2023). Domain Specific Assistant Instruction on Psychotherapy Chatbot. Submitted to Knowledge-Based Systems (Major Revision)

In Chapter 2, we constructed the abnormal brain network connections of depression by using EEGs, and relying on these abnormal connections, we explored the feasibility of utilizing Artificial Intelligence (AI) and EEGs for depression detection, as well as depressive severity classification. In order to prove and verify this method, we respectively collected EEG (52 healthy and 48 depressed participants) from one university and one hospital, now publicly available to researchers in the field. We found that EEG signals extracted from the beta band are more distinctive in depression classification, and these selected channels tend to perform better in scoring depressive severity. This chapter also uncovered the different brain architectural connections by relying on phase coherence analysis. Increased delta deactivation accompanied by strong beta activation is the main feature of depression as the depression becomes more severe. We concluded that the model developed in this chapter is acceptable for classifying depression, as well as for depressive severity. The model can offer physicians a topological dependency, a quantified semantic depressive severity and better models of classifying or scoring depression using EEG signals. Additionally, these selected brain regions and significant beta frequency bands can improve the performance of the BCI system.

In Chapter 3, we propose a FW-GPL for the ordinal scoring task which is also called an ordinal regression problem. Many published conventional methods treat depressive severity estimation as a traditional regression problem and make a strong assumption that each category owns adequate instances to outline its distribution. Our FW-GPL aims to adaptively refine the ordinal label pattern by using two novel techniques: **(1)** assembling fuzzy logic to the fully connected layer of convolution neural networks, and **(2)** transferring the ordinal labels to Gaussian processed labels. Specifically, it incorporates a heuristic fuzzy logic from the ordinal characteristic, simultaneously plugging in ordinal distribution shapes which penalize the difference between the targeted label and its neighbours to ensure a concentrated regional distribution. Accordingly, the function of these proposed windows is leveraged to minimize the influence of majority classes that mislead the prediction of minority samples. Our model is specifically designed to carefully avoid the case of partial missing of continuous ordinal segments. Extensive experimental results on several ambiguous image datasets and one EEG dataset of depression demonstrate that our FW-GPL can achieve compelling performance compared to the State-Of-The-Art (SOTA).

In Chapter 4, we used Phase Lock Coherence (PLC) and General Partial Directed Coherence (GPDC) to construct connections among four adaptively fitted EEG sources, and we also applied previous published models to describe the brain circuits of maintenance, inhibition and disinhibition. Referring to a classical visual n-back paradigm, we recruited forty five mental health undergraduates in this experiment.We found that the bilateral Prefrontal Cortex (PFC) mainly focused on some cognitive components, for example, rehearsal before recognition to classify objects, inhibition to maintain positive memory and activities, and disinhibition to arouse or activate subsequent interactions in brain. Meanwhile, the right PFC sometimes could assist left PFC to implement high capacity WM tasks. By contrast, the posterior regions, Posterior Parietal Cortex (PPC), tends to be engaged in attention arousing and maintaining. These two findings suggest that **a)** the recurrent maintenance circuit may keep the brain executing positive cognitive components, **b)** then the instantly monitoring inhibition would pause the deadlocked sustention function to save energy, and **c)** the arriving of disinhibition arouses the next step in brain to select new subject or focus on novel subjects.

In Chapter 5, we proposed one PEFT method inspired by neuroscience knowledge. Fine-tuning pre-trained LMs may not always be the most practical approach for downstream tasks. While adaptation fine-tuning methods have shown promising results, a clearer explanation of their mechanisms and further inhibition of the transmission of information is needed. To address this, we propose an InA fine-tuning method that aims to reduce the number of added tunable weights and appropriately reweight knowledge derived from pre-trained LMs. The InA method involves **(1)** inserting a small trainable vector into each Transformer attention architecture and **(2)** setting a threshold to directly eliminate irrelevant knowledge. This approach draws inspiration from the shunting inhibition, which allows the inhibition of specific neurons to gate other functional neurons. With the inhibition mechanism, InA achieves competitive or even superior performance compared to other fine-tuning methods on $BERT-large$, $RoBERTa-large$, and $DeBERTa-large$ for text classification and question-answering tasks.

Fine-tuning pre-trained LMs) may not always be the most practical approach for downstream tasks. While adaptation fine-tuning methods have shown promising results, there is a need for a clearer explanation of their mechanisms for approximating the reweighting of pre-trained LMs. To address this, we propose an InA fine-tuning method that aims to reduce the number of added tunable weights and appropriately reweight knowledge derived from pre-trained LMs. The InA method involves **(1)** inserting a small trainable vector into each Transformer attention architecture and **(2)** setting a threshold to directly eliminate irrelevant knowledge. This approach draws inspiration from the threshold used in Spike Neural Network (SNN), which allows the inhibition of specific neurons to gate other functional neurons. With the inhibition mechanism, InA achieves competitive or even superior performance compared to other fine-tuning methods on $BERT_{large}$, $RoBERTa_{large}$, and $DeBERTa_{large}$ for text classification and question-answering tasks.

In Chapter 6, we proposed one instruction tuning method based on the assistant of LLMs. LLMs have demonstrated impressive generalization capabilities through fine-tuning on specific tasks with human-written instruction data. However, the limited quantity, diversity, and professional expertise of such instruction data raise concerns about the performance of LLMs in psychotherapy tasks when provided with domain-specific instructions. To address this, we propose Domain-Specific Assistant Instructions based on AlexanderStreet therapy and counseling data and fine-tune pre-trained LLMs on this dataset. Through quantitative evaluation of linguistic quality using automatic and human evaluation, we observe that pre-trained LLMs fine-tuned on Psychotherapy Assistant Instructions outperform SOTA LLMs response baselines. Our Assistant-Instruction approach offers a half-annotation method to align pre-trained LLMs with instructions. We also release our large synthetic dataset, facilitating future studies on professional instruction tuning.

I end this thesis with an overarching framework of how we can conceptualize depressive severity detection and allusions to possible avenues of helping depression recovery in future research.

Chapter 2

Classifying and Scoring Major Depressive Disorders by Selecting Frequencies and Channels with the Use of Residual Neural Networks

MDD – can be evaluated by advanced neurocomputing and traditional machine learning techniques. This study aims to develop an automatic system based on a BCI to classify and score depressive patients by specific frequency bands and electrodes. In this study, two Residual Neural Networks (ResNets) based on EEG monitoring are presented for classifying depression (classifier) and for scoring depressive severity (regression). Significant frequency bands and specific brain regions are selected to improve the performance of the ResNets. The algorithm, which is estimated by 10-fold cross-validation, attained an average accuracy rate ranging from 0.371 to 0.571 and achieved average RMSE from 7.25 to 8.41. After using the beta frequency band and 16 specific EEG channels, we obtained the best-classifying accuracy at 0.871 and the smallest RMSE at 2.80. It was discovered that signals extracted from the beta band are more distinctive in depression classification, and these selected channels tend to perform better on scoring depressive severity. Our study also uncovered the different brain architectural connections by relying on phase coherence analysis. Increased delta deactivation accompanied by strong beta activation is the main feature of depression when the depression symptom is becoming more severe. We can therefore conclude that the model developed here is acceptable for classifying depression and for scoring depressive severity. Our model can offer physicians a model that consists of topological dependency, quantified semantic depressive symptoms and clinical features by using EEG signals. These selected brain regions and significant beta frequency bands can improve the performance of the BCI system for detecting depression and scoring depressive severity.

2.1 Introduction

MDD, a mental illness, often entails a heightened risk of suicidal thoughts [12]. Individuals grappling with depression frequently face misdiagnosis by physicians, leading to a myriad of issues, including self-medication, substance abuse, inappropriate treatment, social isolation, and impaired academic or work performance [13], [14]. Cognitive behavioral therapy is effective for mild depression, while the combination of psychotherapy and antidepressant drugs is currently the most successful treatment for severe depression [15]–[17]. Inadequate treatments can result in future relapse and prolonged discontinuation symptoms [18].

Depression is commonly categorized as non-depressed, mild, moderate, or severe based on the severity of symptoms [19]. However, a descriptive study revealed a high misdiagnosis rate of MDD, reaching 65.9% [14]. This implies a primary accuracy rate of less than 35% [14]. The failure to accurately diagnose MDD stems from insufficient clinician training and inadequate early-stage appointments, examinations, and treatments for sufferers [14], [20]. Existing diagnostic tools for MDD face challenges, including being time-consuming, requiring trained personnel for administration, inability to classify severity, and lacking visualization results, such as brain topological maps [21], [22].

To address these challenges, we hypothesize that delta and beta brain activities are linked to depression, as indicated by previous studies [1], [7], [8], [71], [72]. In our pursuit of early depression detection, we analyze delta and beta activities and corresponding brain networks, visualizing the results. The Phase Synchrony Index (PSI) [1], [7], [8], [71], [72] is computed to construct brain functional networks, selecting electrodes and frequency bands based on different PSIs between depressive and healthy groups. Subsequently, a classifier utilizing ResNet [73] is designed to process selected EEG signals and detect depression. Additionally, a regression model relying on ResNet is proposed to score depressive severity. Both optimized ResNets on EEGs aim to expedite computation and diagnosis, making this BCI system a complementary tool for depression detection, severity monitoring, and evaluating conventional treatments in healthcare settings.

The contributions of this paper are: (**1**) Presenting central-parietal increased delta deactivation with strong beta activation in the severe depression group during working memory tasks. (**2**) Proposing a classification ResNet with specific frequencies and brain regions for improved and practical depression detection. (**3**) Introducing a regression ResNet with specific frequencies and brain regions for scoring depressive severity based on professional psychologists' labels. The codes and corresponding documentation can be found here: `https://github.com/ChengKang520/Classifying-and-Scoring-MDD-BCI`.

2.2 Related Works

Detecting depression at an early stage is crucial to ensure timely and effective treatment, preventing prolonged suffering and potential suicide. Machine learning approaches have emerged as valuable tools for early detection.

In recent years, advancements in medical imaging and methods utilizing electrophysiological signals have been significant. Many of these approaches focus on extracting brain networks and employing diagnostic models. The experimental flow, as depicted in Figure 1.2, illustrates the integration of these methods. Previous studies, including [1], [7], [8], have highlighted differences in delta and beta brain activities between individuals with depression and those in control groups.

Building upon these insights, we have designed a specialized system, detailed in Figure 1.2. The process, depicted from **A1** to **A4**, begins with **A1**, which involves calculating the PSI between two EEG signals. The formulas for computing PSI can be denoted as follows:

$$\Delta\theta_{trialk}^{n\to m} = \theta_{trialk}^{n} - \theta_{trialk}^{m} \tag{2.1}$$

$$r^{n\to m} = \frac{\sqrt{\left\{\sum_{trialk=1}^{N} sin(\Delta\theta_{trialk}^{n\to m})\right\}^2 + \left\{\sum_{trialk=1}^{N} cos(\Delta\theta_{trialk}^{n\to m})\right\}^2}}{N} \tag{2.2}$$

$$lag^{n\to m} = arctan\{\frac{\sum_{trialk=1}^{N} sin(\Delta\theta_{trialk}^{n\to m})}{\sum_{trialk=1}^{N} cos(\Delta\theta_{trialk}^{n\to m})}\} \tag{2.3}$$

where $\Delta\theta_{trialK}^{n\to m}$ is the difference between the angles of two electrodes ($\Delta\theta_{trialK}^{n}$ and $\Delta\theta_{trialK}^{m}$) under the k-th trial. N is the number of total trials, and $r^{n\to m}$ is the mean value. We also denote it as PSI. Lastly, $lag^{n\to m}$ is the averaged angle of N trials. The entire procedure for constructing brain functional connection networks is presented in our previous study [8]. Moreover, **A2** shows the significant features that we have detected from functional brain networks during working memory tasks - beta frequency band and 16 selected electrodes out of 64. After these pre-processing steps, **A3** shows the use of ResNet architectures [73] and lists out the strategies for classifying and scoring depression. **A4** shows two outputs which consist of the detection result for depression and the score of grading depressive severity.

2.2.1 Brain regions and extraction of functional networks

Methods centered around functional or structural brain networks play a pivotal role in various mental health diagnosis approaches. These techniques are particularly instrumental in identifying conditions such as bipolar disorders [74], [75] and schizophrenia [76], with a specific focus on depression detection through EEG analysis [77]–[79]. To enhance accuracy in detection, researchers have concentrated on extracting pertinent information during the initial pre-processing phase.

In the initial stage, when forming functional brain networks or selecting key brain regions, various indices are computed to estimate interconnections or spectral characteristics among these regions. For instance, a study during the resting state utilized Adaboost classifiers, employing spectral coherence, to identify Cognitive Emotion Regulation Strategys (CERSs) [80]. The distinct advantages of the spectral patterns in the left and right frontal-prefrontal regions were evident in estimating depressive symptoms during the resting state [81]. The absolute power of the theta wave emerged as a reliable characteristic for discriminating depression, leading researchers to employ K-Nearest Neighbor (KNN) with 10-fold cross-validation for classification [77]. Subsequent to calculating relative wavelet energy and various entropy features via Decomposed Discrete Wavelet Transform (DWT) coefficients on EEG signals, a feed-forward ANN was employed for depression classification [79]. A feature-level fusion approach was adopted to identify robust features, and traditional machine learning classifiers were then applied for depression detection using multimodal EEG data [78].

Brain networks exhibit recognizable cognitive patterns, such as the abnormal cognitive control network observed in depressive patients [82]. Additionally, these networks reveal electrophysiological connections in various frequency bands (delta, theta, alpha, and beta) [8]. Leveraging brain oscillations in different frequencies, the PSI [1], [7], [8], [71], [72] is calculated to construct functional brain networks. PSI reflects the degree of synchronization between two EEG channels, and subsequent correlation coefficient calculations based on PSI facilitate the use of an online clustering approach to construct convergent brain networks, as detailed in prior studies [1], [7], [8], [71], [72]. Consequently, Morlet's wavelet is employed to calculate the time-frequency domain and the corresponding angle:

$$\varphi^{n}_{trialk}(f,t) = \frac{1}{\sqrt{\pi}\delta_t} exp(\frac{-t^2}{2\delta_t^2}) exp(j2\pi f t) \tag{2.4}$$

$$\Delta\theta^{n\to m}_{trialk} = angle\{exp(i[\varphi^{n}_{trialk}(f,t)])\} - angle\{exp(i[\varphi^{m}_{trialk}(f,t)])\} \tag{2.5}$$

where $\varphi^{i}_{trialk}(f,t)$ is the Morlet's wavelet at frequency domain f, and δ_t is the standard deviation of the Gaussian function $\varphi^{n}_{trialk}(f,t)$. When relying on the EEGLAB in the

MATLAB environment, the wavelet cycles and the lowest time-frequency window are selected referring to our previous studies [1], [7], [8].

2.2.2 Artificial neural networks utilization

Classifying depression through machine learning approaches significantly contributes to expediting the diagnostic process. Leveraging machine learning methods such as (Support Vector Machine (SVM), AdaBoost, and Random Forest (RF), the most widely employed clinical techniques involve Magnetic Resonance Imagings (MRIs) and EEGs. To enhance efficiency, only carefully selected channels are utilized during training tasks to prevent information overload, which may otherwise impede training and lead to model overfitting [83]. The typical workflow of depression detection systems can be delineated into three distinct steps.

Step 1: Psychological Paradigm. Research indicates that adaptive dual n-back WM training can alleviate subclinical anxiety and depression symptoms in adolescents [84]. The learning processes of WM capacity were found to moderate the relationship between Brain Derived Neurotrophic Factor (BDNF) and psychotherapy outcomes for depression [85]. This highlights the crucial role of WM in reflecting depression severity. The n-back paradigm was chosen for two reasons: **(1)** to control emotional task intensity, a designed n-back paradigm evaluates participants' WM capacity by adjusting task difficulty (ranging from 0-back as the baseline to higher levels); **(2)** to explore potential rehabilitation methods through WM training for future endeavors.

Step 2: Feature Extraction. Neuroimaging regions [82] and electrophysiological areas [19] are commonly extracted and fed into machine learning models. Additionally, selected EEG channels contribute to depression classification during resting or task-completion states with eyes closed.

Step 3: Classification and Scoring of Depressive Symptoms. In recent years, traditional machine learning approaches with EEG have been employed to identify depressed subjects using methods such as ANN [79], logistic regression [86], SVM [87], bagged tree [88], and Convolution Neural Network (CNN) [19]. Deep learning methods, especially CNN architectures, automatically extract crucial features and score depression severity across various psychological tasks. However,Long Short Term Memory (LSTM), a time-series model, necessitates participants to complete long continuous tasks.

For scoring the severity of depressive symptoms, one study utilized fMRI images and a kernel partial least squares regression model, evaluating performance with RMSE [89]. Finally, our proposed approach, based on beta EEG and sixteen specifically selected channels, employs ResNets due to their ability to mitigate gradient vanishing and enable deeper architectures with fewer parameters [73].

2.3 Methodology

2.3.1 Participants and EEGs Recording

The EEG signals utilized in this study were sourced from Shenzhen University and Shenzhen Kangning Hospital in Shenzhen, China, with approval from the ethics committee of Shenzhen Mental Health Center. The dataset comprises 52 healthy undergraduate dextromanual students (with a gender distribution of 6:4 males to females and a mean age of 20.4 ± 9.7) and 48 depressed patients (with a gender distribution of 6:4 males to females and a mean age of 34.3 ± 12.1). Rigorous selection and assessment procedures, as detailed elsewhere [1], were employed. In both the healthy and depressive groups, participants had not taken any medication, and there was no personal or family history of psychiatric or neurological diseases before the experiments. Depressed scores were assessed using the Structured Clinical Interview for DSM-IV (SCID-CV) [23] and the 17-item Hamilton Rating Scale for Depression (HAMD), administered by two professional clinical psychologists. Depressive patients were screened before the experiment, and those with discordant scores (one mild, one severe) were excluded. Conversely, the final score label was determined by computing the average score if the two psychologists' scores were at the same level. Depressive severity, based on the HAMD and SCID-CV, was classified, and the 17-item HAMD cut-off points were defined as follows: severe ($\geq$24), moderate (17-23), mild (8-16), and non-depressed (0-7) [38]. Due to the small difference between moderate and mild depression and an imbalance in the EEG data distribution after preprocessing, we consolidated the categories into three groups to mitigate potential risks. Consequently, the selected groups in this system are healthy controls (non-depressed: 0-7), depressed with low scores (Score: 8-23), and depressed with high scores (Score: $\geq$24).

2.3.2 Working Memory Experiments

Building upon our prior investigations [1], [8], the n-back experiment discussed here was conducted within the E-Prime 5.0 environment. We employed the letter variant of the n-back tasks, designating 0-back tasks as the baseline and 1-back and 2-back tasks as the WM load. Participants were tasked with observing and responding to black letter stimuli presented on a white background. Simultaneously, they were required to press specific buttons—using the index finger for matching stimuli and the middle finger for mismatching stimuli. In the 0-back tasks, participants identified a pre-specified letter 'X' on the screen by pressing the matching button. In the 2-back tasks, they pressed the matching button if they recognized a letter that corresponded to the one presented two trials before. The presented letters were randomly selected from English consonants.

The experiment was divided into three segments, each containing three tasks (0-back, 1-back, and 2-back) arranged in a random sequence to prevent any potential performance bias caused by a predictable and fixed sequence design. Task durations were set at 75 seconds for each segment, comprising a pseudo-random sequence of 30 consonants (10 targets and 20 non-targets). To ensure accurate manipulation and allow sufficient reaction time, letters were presented for 0.5 seconds and disappeared in the subsequent 2 seconds. Participants were provided with 45 seconds between every two segments for breaks. Behavioral performance, including reaction time and response accuracy rate, was recorded. Notably, incorrect responses were excluded during the electroencephalogram (EEG) analysis. Once participants clarified any queries and confirmed their understanding of the details, the warm-up tasks, serving as guidance before the formal experiment, were concluded.

2.3.3 Preprocessing of EEGs before Training

All procedures, encompassing EEG recording and preprocessing, have been meticulously detailed in previous studies [1]. In summary, the following steps were undertaken: (**1**) removal of eye movements, (**2**) band-pass filtering within the frequency range of $0.16 \sim 30$ Hz (24dB/Octave), (**3**) artifact rejection, and (**4**) baseline correction. Notably, the computation of phase coherence was executed prior to the training tasks. This preprocessing is integral to the overarching goal of this study, which aims to develop an automated system capable of classifying depression and scoring depressive severity based on selected frequency bands and electrodes. Brain connection maps were constructed utilizing the phase coherence method, a technique detailed in our prior works [1], [7], [8], [71], [72]. The input data for this process comprises EEG signals collected from either 64 or 16 channels during three distinct task types: 0-back, 1-back, and 2-back. This comprehensive approach lays the foundation for the development of a robust system for the classification and severity assessment of depression.

2.3.4 Residual neural networks

In Figure 2.1, 64 channels recorded EEG signals over a duration of 2.5 seconds. Subsequently, a down-sampling process reduced the data length from 2500 points to 1250 points. After discarding 98 points from the tail, the input size for the first model was set as $64 \times 64 \times 18$. Two residual neural networks were employed to train the EEG data for 0-back, 1-back, and 2-back tasks. In the second training phase, 16 electrodes selected using the phase synchronization method resulted in an input size of $16 \times 64 \times 18$. The total size of the EEG data amounted to 22.5 million sampling points (48 depressive patients +

52 healthy controls) * 60 trials * 3 tasks (0-back, 1-back, and 2-back) * 2.5 seconds * 500 sampling rates = 22.5 million. Testing the CNN with 6 residual blocks yielded optimal performance, with a parameter size of 0.85 million, effectively preventing overfitting or underfitting issues through proper parameter selection.

Given the widely recognized 65.9% misdiagnosis rate of MDD [14], we set the detection rate threshold at 70%. Each participant underwent 60 trials, and the depressive probability for a participant was determined by the ratio of trials with predicted probabilities exceeding 70% to the total number of trials. If the predicted probability for a subject on a trial exceeded 70%, the system classified them as 100% depressive for that trial. Finally, if, during a trial, 33 out of 40 subjects had probabilities from the ResNet classifier equal to or greater than 70%, the model's accuracy rate was 82.5% (33/40). Additionally, the second ResNet regression model provided the severity score of depression, referencing the SCID-CV system and the HAMD score.

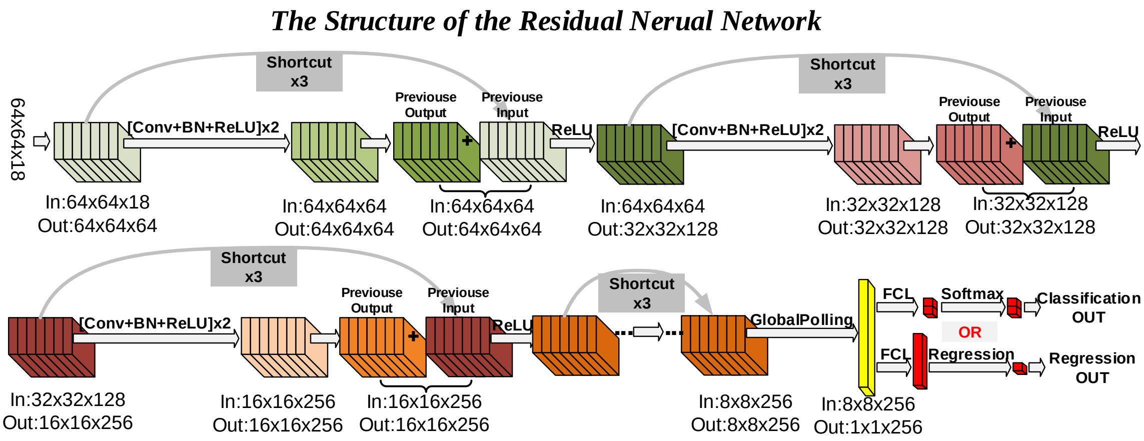


Figure 2.1: The structure of the constructed residual neural network. The input size is $64\times64\times18$ or $16\times64\times18$. Conv+BN+ReLu means the processing of convolution (Conv), batch normalization (BN) and rectified linear unit (ReLU). FCL is the fully connected layer. The shortcut is purely forward plus. $\times3$ means this block should be repeated triple times.

2.4 Result

2.4.1 Memory load comparison of behavioural results

Table 2.1 shows the significant level between the low and the high depressed group in terms of response accuracy rate and reaction time during three different working memory tasks (0-back, 1-back and 2-back). During the 0-back task, there is no significant difference ($P = 0.061$) between MDDs with low scores and the MDDs with high scores in terms of the response accuracy rate. But for the reaction time, the difference is significant

Table 2.1: The comparison of reaction time and response accuracy rates between two different memory loads (average ± standard deviation) in two depressive groups.

Memory Load	0-back		1-back		2-back	
Scores	**Low scores**	**High scores**	**Low scores**	**High scores**	**Low scores**	**High scores**
Response Accuracy	98.9±1.4	96.3±2.2	92.8±4.7	86.3±3.6	84.9±5.3	75.5±7.6
Reaction Time	545±53	561±47	701±147	751±129	769±176	791±183
Statistics	**P value**		**P value**		**P value**	
(the low and the high)	Accuracy Rate	Reaction Time	Accuracy Rate	Reaction Time	Accuracy Rate	Reaction Time
	P = 0.061	P = 0.017	P <0.01	P <0.01	P <0.01	P = 0.053

($P = 0.017$). In the 1-back task, both the response accuracy rate and the reaction time show a significant level ($P < 0.01$). When implementing the 2-back task, the MDDs with low scores demonstrated a significant difference in response accuracy rate ($P < 0.01$).

2.4.2 The Connections comparison

We designated the 0-back task as the "rest-state" and the 2-back task as the cognitive load for WM. Consequently, a decrease in PSI signifies diminished neuronal activity in the corresponding brain regions, indicating a return to the 'rest-state.' Conversely, an increase in PSI corresponds to heightened neuronal activity in these regions, reflecting robust WM-related mechanisms.

Figure 2.2 depicts the number of significantly connected pairs based on the two WM tasks (0-back and 2-back). Notably, for PSI decrease, the most significant differences among the three groups lie in connections within the entire delta frequency components. The high-scoring depressed group predominantly exhibits whole theta frequency connections, while other frequency bands show no significant differences. In contrast, considering PSI increase, the high-scoring depressed group displays the fewest delta, theta, and alpha-connected pairs. Both depressed groups exhibit a higher number of whole beta connections, with the low-scoring depressed group demonstrating stronger connections in the delta, theta, and alpha bands compared to the high-scoring depressed group. To gauge the overall impact (the product of the number of significant pairs and the corresponding PSI values), representing significant PSI levels across the three groups, t-values resulting from a two-sample t-test are shown in Figure 2.3 ($P < 0.01$). The most significant frequency component is highlighted in each histogram. Aside from Figure 2.3 B, which indicates a slightly pronounced frequency part in the delta bands ($P < 0.05$), the last three histograms underscore that beta frequency activation constitutes the most significant difference.

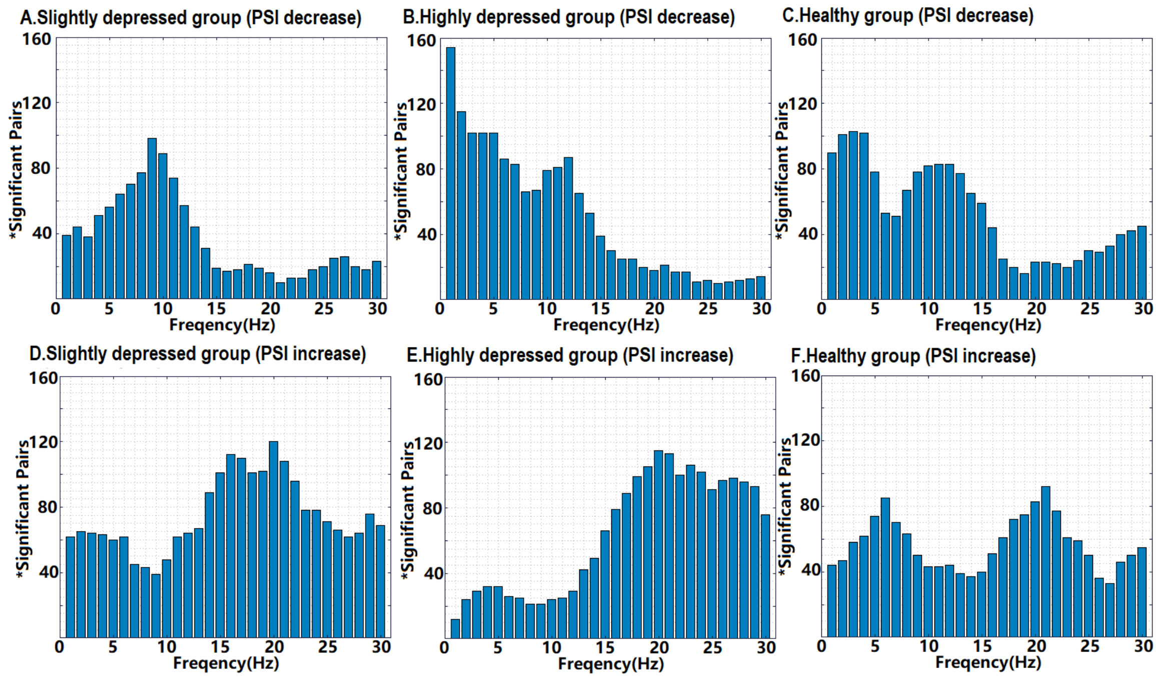


Figure 2.2: The number of the significant pairs in terms of the comparison between 2-back and 0-back tasks.

2.4.3 Clusters between these Three Groups

In accordance with the comparison of PSI connections, depicted in Figure 2.4A, a discernible PSI decrease is observed in the depressive group with lower scores, contributing to a sparse electrode connection pattern. This decrease is further illustrated by the flat distribution in the beta frequency band, as depicted in Figure 2.2. Conversely, the PSI increase shown in Figure 2.4B indicates that the control group fails to generate a cohesive cluster. In contrast, the depressive group with low scores tends to concentrate connected pairs in the left parietal and left central regions, forming what is referred to as Cluster A.

In the comparison between the depressive group with high scores and the control group, as illustrated in the lower panel of Figure 2.4C, the PSI decrease reveals that the control group has fewer connected pairs, primarily in the left frontal and whole parietal areas (Cluster C). Meanwhile, the depressive group demonstrates nearly comprehensive cerebral connections, excluding the occipital areas (Cluster B). Conversely, concerning the PSI increase, the depressive group with high scores exhibits a compact connecting pattern involving the left frontal-central and right central-parietal regions. Additionally, there is connectivity observed in the left frontal-temporal and right temporal-parietal areas, as depicted in Cluster D.

Table 2.2: Classification (Accuracy) and scoring depression (RMSE) results (mean and standard deviation) using whole frequency bands (delta, theta, alpha and beta).

Accuracy Rate (>0.70)	0-back	1-back	2-back	Best Result
	0.457±0.063	0.429±0.100	0.514±0.164	0.734 in 2-back
Score Difference (RMSE)	0-back	1-back	2-back	Best Result
	8.38±3.22	8.41±3.52	7.73±3.22	3.22 in 2-back

Table 2.3: Classification (Accuracy) and scoring depression (RMSE) results (mean and standard deviation) using beta frequency bands.

Accuracy Rate (>0.70)	0-back	1-back	2-back	Best Result
	0.514±0.217	0.429±0.226	0.371±0.217	0.783 in 0-back
Score Difference (RMSE)	0-back	1-back	2-back	Best Result
	7.97±2.25	7.59±1.51	8.05±1.40	4.10 in 0-back

Table 2.4: Classification (Accuracy) and scoring depression (RMSE) results (mean and standard deviation) using whole frequency bands (delta, theta, alpha and beta) and selected EEG channels.

Accuracy Rate (>0.70)	0-back	1-back	2-back	Best
	0.514±0.217	0.514±0.239	0.571±0.141	0.714 in 1-backs
Score Difference (RMSE)	0-back	1-back	2-back	Best
	7.77±3.11	7.25±2.19	7.37±2.14	2.88 in 1-back

Table 2.5: By scaling the size of proposed ResNets, the below shows the classification (Accuracy) and scoring (RMSE) results using beta frequency band and selected EEG channels.

ResNet (Size: 2.4M)				
Accuracy Rate (>0.70)	0-back	1-back	2-back	Best
	0.452±0.302	0.409±0.222	0.414±0.367	0.833 (0-back)
Score Difference (RMSE)	0-back	1-back	2-back	Best
	8.12±3.38	8.07±3.44	7.74±3.66	3.02 (0-back)
Default ResNet (Size: 4.6M)				
Accuracy Rate (>0.70)	0-back	1-back	2-back	Best
	0.429±0.226	0.514±0.126	0.457±0.234	0.871 (2-back)
Score Difference (RMSE)	0-back	1-back	2-back	Best
	7.97±3.57	7.83±3.31	7.59±3.83	2.80 (2-back)

2.4.4 The result of classifying and scoring MDD patients

Following the aforementioned pre-processing steps, each subject underwent no more than 60 trials due to the exclusion of substandard trials. The outcomes encompassing the entire frequency band are detailed in Table 2.2, while specific results for the beta frequency bands are presented in Table 2.3. In the second model, an expansion of the system was undertaken to classify depression and assess depressive severity by extracting the beta frequency band and identifying 16 significant electrodes. The online clustering step, utilizing PSIs, produced Cluster A and Cluster D. The electrodes—Fz, F1, F3, FCz, FC1, FC3, FC5, FT7, FT9, T7, CP3, CP2, CP4, CP6, TP8, and TP10—most frequently connected in both Cluster A and Cluster D, significantly contributed to enhancing the performance of depression classification and depressive severity scoring. To mitigate randomness in results, a 10-fold computing method was employed to select the optimal outcome. For instance, Table 2.4 illustrates a classification accuracy rate of 0.714 when utilizing the entire frequency bands. However, Table 2.5 indicates that relying on the beta frequency bands can yield an accuracy rate as high as 0.871. Ultimately, in the context of 2-back tasks within beta frequency bands and with the contribution of specifically selected channels, the maximum accuracy achieved through 10-fold testing is 0.871. Concerning the assessment of depressive severity, despite achieving a minimum RMSE result of 2.8 in 2-back tasks with beta frequency bands and specifically selected channels (as shown in Table 2.5), the overall performance in scoring depressive severity (Table 2.5) is inferior to that in Table 2.4.

2.5 Discussion

In this study, deactivation refers to the dominance of the rest state, while activation signifies the processing of working memory. Our observations reveal that the low depressive group exhibits weaker delta deactivations but stronger beta activations. Conversely, the high depressive group displays more pronouncedly deactivated delta connections and heightened activation of beta connections. Furthermore, as depressive severity intensifies, beta right central parietal functional connections emerge in depression patients. Moreover, the beta frequency bands play a significant role in distinguishing depressive patients from healthy controls. Specific channels, when selectively chosen, prove effective in easily differentiating depressive patients. Utilizing beta frequency bands enhances the accuracy of scoring depressive severity, with the selected channels exhibiting notable scoring advantages within the beta frequency band.

2.5.1 Possible inducing reason for getting depression

As depressive symptoms intensify, individuals with depression exhibit more pronounced delta deactivations and beta activations. Interestingly, there is no conclusive evidence supporting the presence of obvious theta and alpha activities in these cases. It is noteworthy that individuals infected with Human Herpesvirus 6 (HHV-6) do not demonstrate a correlation with theta and alpha EEG oscillations, as reported by [90]. Furthermore, for patients afflicted with HHV-6 infection, subsequent to medical intervention and a 14-day recovery period, there is a noticeable slowing down of theta/delta EEG oscillations [91]. This deceleration implies a weakening of theta/delta activities [92]. Notably, Human Betaherpesvirus 6B (HHV-6B) infection has been associated with an increased potential risk of mental disorders, particularly depression [93], [94]. In light of these findings, we can infer a potential connection between HHV-6 and depression. Our subsequent research will delve into the extent to which HHV-6 may contribute to the development of depression.

2.5.2 Topological analysis

The topological networks approach facilitates the comparison of distinct cognitive patterns. On one hand, the analysis of phase coherence reveals that individuals in the depressive group tend to exhibit diminished low-frequency WM activation, particularly in the delta and theta frequency bands. This trend becomes more pronounced as depressive symptoms progress from moderate to severe. On the other hand, as depicted in Figure 2.3 C and D, highly depressed patients display a significant disparity in beta WM activation compared to those in the mildly depressed group. Consequently, the depressive group exhibits stronger beta activations than healthy controls, with highly depressive patients being at a heightened risk of experiencing this imbalance. Furthermore, mildly depressed patients demonstrate a deficiency in delta and theta WM deactivation, whereas the highly depressive group exhibits redundant delta and theta WM deactivation. During WM tasks, the depressive group reports reduced frontal-midline theta power and increased occipital upper alpha power during WM encoding [95]. This aligns with similar research suggesting that depressive patients manifest abnormal brain activities across all frequency bands. The topological structure of beta frequencies (Cluster D in Figure 2.4) among highly depressive patients reveals additional central-parietal WM activation compared to that of slightly depressive patients (Cluster A in Figure 2.4). This corresponds to findings indicating that MDD is characterized by unique EEG oscillations in beta frequencies, which dominate over delta, theta, and alpha when compared with healthy subjects [96], [97]. High beta coherence is also associated with connections within and between the Dorsolateral Prefrontal Cortex (DLPFC) or temporal regions [81].

The increased delta deactivation during WM tasks signifies low WM loads and may be linked to the resting recovery mechanism from cognitive maintenance. The comparison between Cluster B and Cluster C (Figure 2.4, panel C), along with the observed increase in the delta band (Figure 2.2), suggests a consistent upward trend in delta deactivation as depressive symptoms intensify, as observed in a neuromodulation therapy study [98]. slightly depressive patients present a lack of delta and theta WM deactivation, and on the contrary, the highly depressive group shows redundant delta and theta WM deactivation. Another study [98] found that increases in beta and gamma power at the Left DLPFC (L-DLPFC) correlated with an improvement in depressive symptoms. Enhanced attentional processes associated with beta and gamma oscillations [99] may explain how Cluster A and Cluster D in beta oscillations could modulate the attention processing of depressive subjects. Comparing Figure 2.2D with the decreased pairs of alpha activation in Figure 2.2E provides similar evidence. Greater reductions in upper alpha and gamma power during WM maintenance are indicative of high depressive severity [95].

2.5.3 Contribution of frequency and topological selection for classifying and scoring depressive patients

In this study, the utilization of a ResNet classifier to differentiate between depressive patients and healthy controls revealed a noteworthy strategy. Specifically, relying on the single beta frequency band proved to yield a higher accuracy rate compared to incorporating all four frequency bands. Moreover, when assessing depressive severity within the depression group, the system introduced an effective method for quantifying the degree of depressive severity. This suggests that the beta frequency holds promise for identifying depression patients during WM tasks [8]. Although beta frequency cerebral activities can serve as a diagnostic tool for detecting depression, they do not contribute to enhancing the accuracy of scoring depressive severity.

It is important to note, however, that within the beta band, the scoring results exhibit wider variances. To ensure more robust assessments of depressive severity, consideration should extend to all frequency bands. The relatively lower average accuracy rate may be attributed to the limited number of psychologists involved in diagnosing patients and providing results—only two in this case. This introduces data instability, particularly when testing deep learning models with potentially misdiagnosed subjects.

2.5.4 State of the art for classifying depressive patients

Table 2.6 highlights the significant advantages of our proposed method, with the highest accuracy rate for detecting depression reaching 87.1%. However, when assessing the

Table 2.6: Comparison with existing methods on classifying depression with EEGs.

References	Subjects	Cross validation	Method + Feature	Accuracy
		EEGs (Scenario)		
Hanshu Cai et al (2020)[78]	MDD = 86, HC = 92	10-fold	KNN + EEGs (Fp1, Fpz, Fp2)	Highest at 86.98%
Xiaowei Zhang et al (2020)[100]	MDD = 81, HC = 89	10-fold	CNN + EEGs + demographic	Average at 75.29%
Xiaowei Li et al (2019)[101]	MDD = 24, HC = 24	24-fold	CNN + EEGs (all frequencies)	80.74% for mild
The proposed method	MDD = 48, HC = 52	10-fold	ResNet + EEGs (beta bands 16 electrodes)	Max: 87.1% and Average at 45.7%

Table 2.7: Comparison with existing methods on scoring depressive severities with EEGs.

References	Subjects	Cross validation	Method + Feature	RMSE
		Images (Scenario)		
Kosuke Yoshida et al (2017)[89]	MDD = 58, HC = 65	leave-one-out	PLS + sMRI	9.56
Benson et al (2012)[103]	MDD = 30, HC = 0	leave-one-out	RVR + MRI	2.50
		EEGs (Scenario)		
The proposed method	MDD = 48, HC = 52	10-fold	ResNet + EEGs (beta bands 16 electrodes)	2.80

overall performance of scoring depressive severity in 2.5, it is observed to be weaker compared to the approach in 2.4, which utilizes the entire frequency bands. We attribute this discrepancy to the potential influence of data quality and the limited robustness of the proposed model. Concerning average accuracy rates, a notable limitation of our proposed method is its inability to consistently yield stable results. Additionally, this approach relies on psychological paradigms, specifically the n-back task, which only captures the brain function associated with working memory.

2.5.5 State of the art for scoring depressive severities

Scoring of depressive severity is addressed in two studies based on MRI-related images with Partial Least Squares Regression (PLSR) and Relevance Vector Regression (RVR) [102]. Table 2.7 shows that under the leave-one-out cross-validation, the minimum RMSE can reach 2.50 [103], which means the RVR+MRI method can precisely grade the depressive severity within 2.50 error. In this study, the proposed method shows a minimum RMSE of 2.80 under 10-fold cross-validation.

2.6 Conclusion and future work

In this study, we introduced a BCI system comprising two models based on the ResNet architecture. The first model aims to detect depression, while the second model focuses on scoring the severity of depressive symptoms. This is achieved using 16 carefully selected channels and beta frequency EEG signals. The ResNet classifier primarily distinguishes depressive subjects from healthy controls, while the ResNet regression model grades the severity of depression. Our study leverages coherence analysis to identify significant frequency bands and brain functional networks in depressive patients, highlighting the role of beta frequency in detecting depression and assessing its severity. The selected EEG channels demonstrate a substantial advantage in classifying depression.

Future research will concentrate on (1) advancing the construction of ANNs, (2) refining EEG data acquisition and selecting participants with depression, (3) designing more robust experiments, and (4) investigating the impact of antidepressant drug treatment. Additionally, we will explore potential connections between the inducing factors of depression and HHV-6.

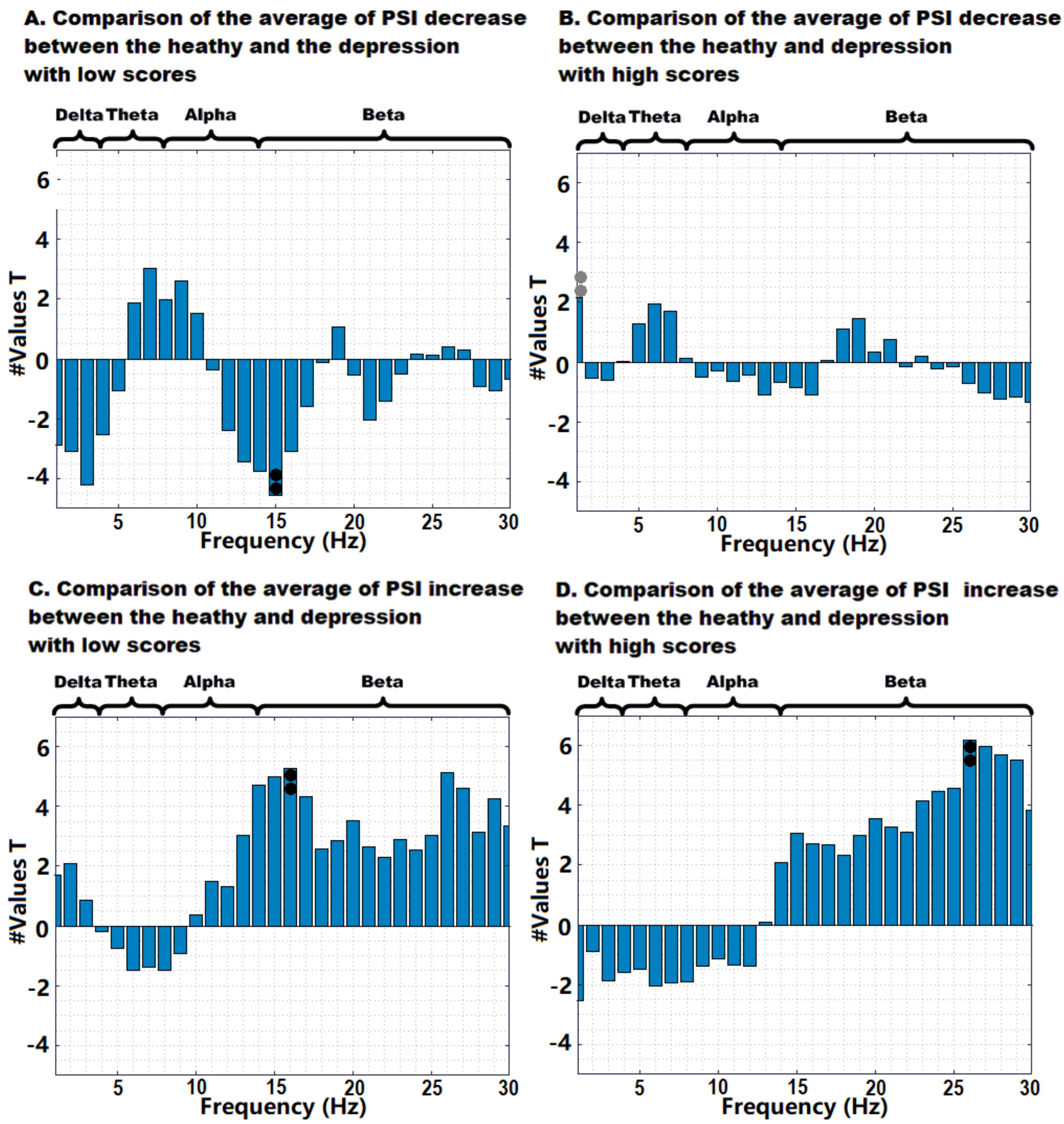


Figure 2.3: The t values (significant level) of the comparison between the depression group and the healthy control group.

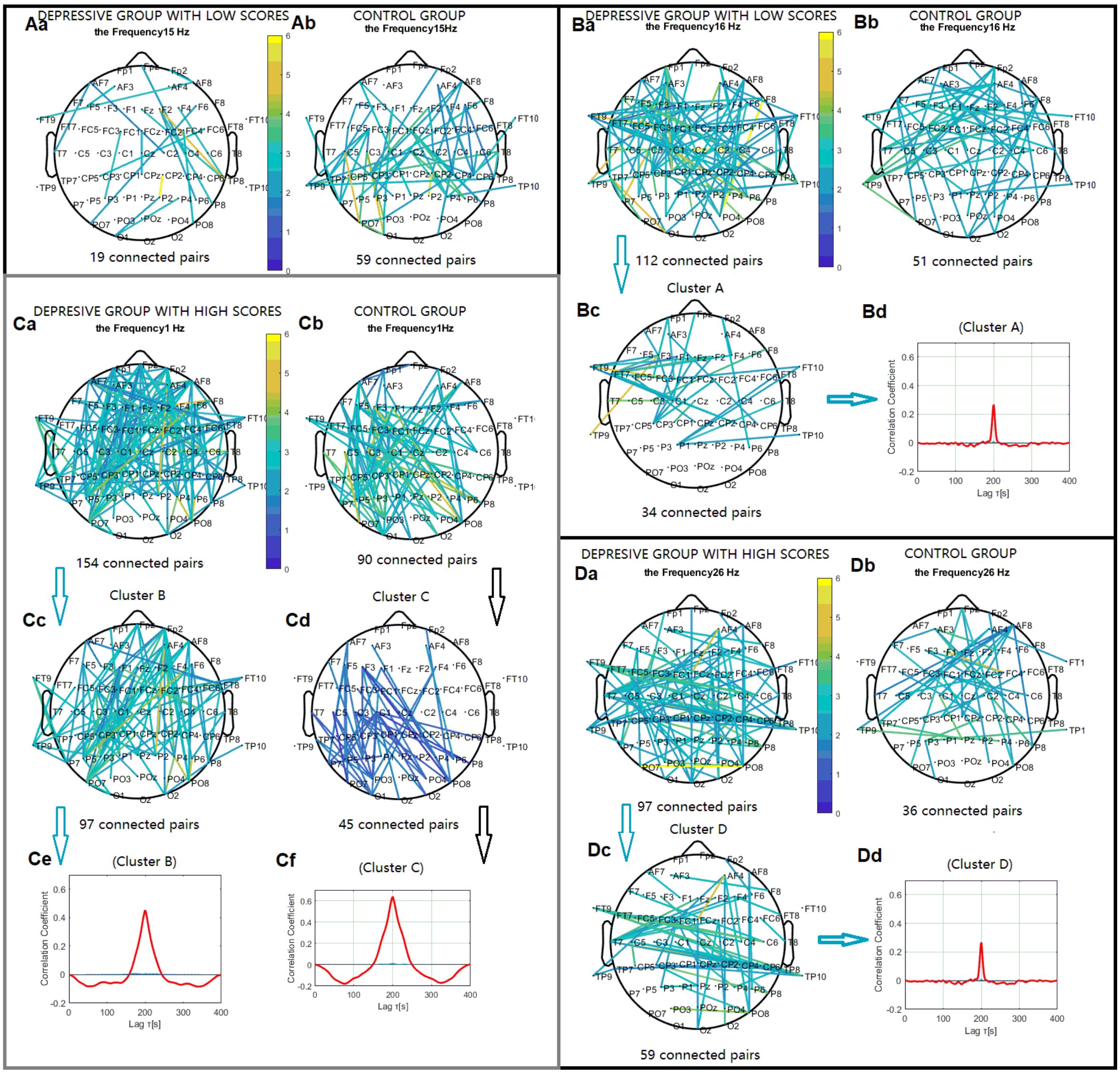


Figure 2.4: Clustering of some significantly increased and decreased phase synchronization indices mainly in beta bands for both the two depression groups and control groups. Lines in the up panel (panel A and B) respectively represent the significant PSI decrease and increase during the 2-back condition. Relative to that during the 0-back condition ($p < 0.05$) between the depressed group with low scores and the control group. Connections in the down panel (panel C and D) respectively represent significant PSI decrease and increase between the depressed group with high scores and the control group. (Bc, Cc, Cd and Dc) Cluster A, B, C and D identified in the control group and two depressed groups respectively were significant using a control of family-wise error rate at the level of $\alpha = 0.01$. Bd, Ce, Cf and Dd are correlation coefficient of phase synchronization within corresponding clusters. The gray panel C means that the significant level is slightly obvious.

Chapter 3

Fuzzy Windows with Gaussian Processed Labels for Ordinal Scoring Tasks

In this chapter, we introduce a FW-GPL designed for unconstrained facial age estimation, framed as an ordinal regression problem. Unlike many conventional methods that treat age estimation as a standard regression problem, assuming each age dataset has sufficient instances to represent its distribution, our FW-GPL seeks to dynamically refine the age label pattern using two innovative techniques: **(1)** incorporating fuzzy logic into the fully connected layer of convolutional neural networks and **(2)** transforming ordinal labels into Gaussian processed labels. Specifically, our approach integrates heuristic fuzzy logic derived from ordinal characteristics, simultaneously incorporating age distribution shapes. This integration penalizes differences between the target label and its neighbors, promoting a concentrated regional distribution. The function of these proposed windows is harnessed to minimize the influence of majority classes that might mislead predictions for minority samples. Our model is explicitly crafted to avoid issues of partial missing of continuous facial age segments, ensuring consistent performance across the entire continuous facial age dataset. Extensive experimental results on three facial aging datasets and one ambiguous medical dataset showcase that our FW-GPL attains compelling performance compared to the SOTA.

3.1 Introduction

Ordinal regression, commonly known as ordinal classification, is a supervised learning problem designed to predict a discrete set of ordinal labels. Its key distinction from traditional classification tasks lies in the fact that the categories bear a natural or implied order.

For instance, in apparent age group estimation, face images are graded on an ordinal scale ranging from "Infants" to "Aged." Ordinal regression is essentially a specialized form of metric regression, where the regression targets are finite and discrete, and the differences in features between adjacent labels may not be uniformly equal. An illustrative example emphasizes that the dissimilarity in facial features between "Infants" and "Children" is more pronounced than that between "Young adults" and "Adults." Ignoring the ordinal relationship reduces the ordinal regression problem to a simple multi-class classification issue.

When dealing with long-tailed age data, a common challenge arises, with head classes tending to dominate training convergence. Consequently, the resulting age classification model performs admirably on head classes but experiences performance degradation in tail classes. This motivates our pursuit of a robust facial age classification approach tailored to imbalanced ordinal data.

Facial age classification approaches can be broadly categorized into two aspects: SLL-Loss functions [40]–[43], [104] and LDBL [40], [44]–[48], [104], [105]. SLL-Loss methods typically process individual facial images independently, disregarding the gradual changes in human facial appearance with aging. This often leads to indistinct or ambiguous facial representations for adjacent age classes. On the other hand, LDBL methods aim to map ordinal ground-truth learning using a Gaussian or Gaussian-like label distribution. However, their performance is challenged in long-tailed scenarios, where the features representing ordinal neighbors or the distinctive features of minority classes are overshadowed by the majority classes.

To address the challenges of long-tailed and conjugated ordinal data, we propose the FW-GPL approach for ordinal regression. The primary objective is to stretch semantic margins, enlarging inter-class variance, particularly for classes sharing common features. As depicted in Figure 1.3, we assume that ordinal neighbor classes have a closely shared feature region, introducing complexity to the ordinal classification task. A fuzzy window with Gaussian-processed labels is meticulously designed atop deep neural networks to mitigate the impact of semantic scoring traps, preserving age-distribution information. Our FW-GPL consists of two crucial components: a difuzzifier window and a learning strategy for Gaussian-processed labels. The difuzzifier window aims to reduce ambiguity in ordinal neighbors while preserving internal ordinal age features. Simultaneously, it seeks to narrow classifier decision boundaries for tail classes by transferring knowledge from head classes.

Practically, Gaussian-processed labels enable the incorporation of prior knowledge (age Gaussian-like distribution), emphasizing major age classes while mitigating the influence of distant neighbor classes. To validate the effectiveness of our proposed method, we conduct extensive experiments on three widely-used face aging datasets: Craniofacial Longitudinal Morphological Face Database (MORPH) II [106], Face and Gesture Recognition Research Network (FG-NET) [107], and Cross-Age Celebrity Dataset (CACD) [108], as well as a medical ordinal dataset, CBIS-DDSM. Our method demonstrates competitive performance, especially in handling fragmentary samples, as the choice of window length effectively trims the influence of tail classes. The main contributions of this work can be summarized as follows:

1. This paper proposes a novel FW-GPL method for facial age estimation. FW-GPL can effectively model the correlation between adjacent ordinal ages and better approximate the age label distribution by avoiding long tails.

2. We also demonstrate that for the age estimation, especially when the age order is not consecutive, FW-GPL can achieve an equivalent level with the wholly sequential age order by selecting a proper length of the fuzzy window.

3. Extensive experiments on FG-NET, MORPH II and CACD datasets show the superiority of our proposed approach to most existing SOTA methods.

3.2 Related Work

The objective of this learning architecture for the ordinal regression problem is to weaken the influence of the overlapping features $F = \{f_1, f_2, \ldots, f_\epsilon\}$ extracted from the neighboring ordinal categories: $C = \{C_1, C_2, ..., C_i, \ldots, C_K\}$ (ϵ is the number of quantized features, and K is the number of categories). Each C_i is an ordinal category containing overlapping features with its neighbors, $\{C_{i\text{-}a}, ..., C_{i\text{-}a\,+\,1}, C_{i\text{-}1}\}$ and $\{C_{i\,+\,1}, ..., C_{i\,+\,b\text{-}1}, C_{i\,+\,b}\}$, where values a and b are related to the relationship between the feature strength of the specific category C_i and its closeness to neighboring categories work [2], [109], the boundary of the window is $\{a, b\}$. In this paper, we also set the upper bound of the window as a and the lower bound of the window as b). Moreover, Gaussian processed labels can prevent extracted features from roughly slipping into one category, which means they make neighboring ordinal categories meticulously divided up according to shared overlapping features.

3.2.1 Ordinal Classification

In the machine learning field, ordinal classification models are reassembled by reformulating the problem to utilize multiple binary classifiers [110]. There are some earlier studies working on constructing CNNs [111], [112], which have replaced the last layer of the ordinal classification model with a number of binary classifiers [113]. In this Ordinal Regression CNN (OR-CNN) architecture, the ordinal classification problem has been converted to a number of K binary classification tasks. If the maximum value of the ordinal label is K, we rearrange the labels with a set $k = \{0, 1, \ldots, K-1\}$ and define the binary classifier as whether the output is greater than k or not. All K binary tasks share the same intermediate layers, but they are assigned distinct weight parameters in the output layer [114]. This OR-CNN architecture deeply relies on the ordinal continuity of the data. If the training dataset has insufficient and intermittent input ordinal labels, and if the dataset has missing data (for example, 150-year-old facial-age data), the fitted OR-CNN cannot recognize the intermittent or missing segment, which inevitably leads to a classification crash.

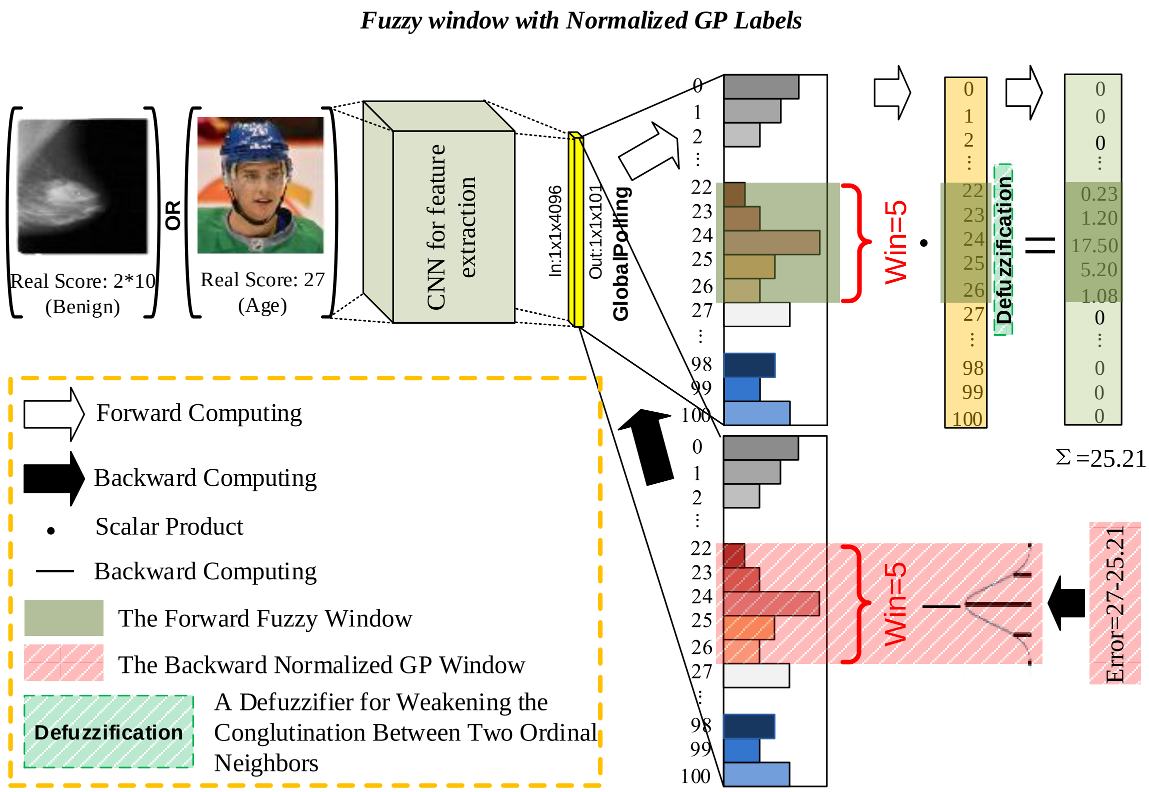


Figure 3.1: The proposed fuzzy window method (the length of the fuzzy window is 5) with the use of Gaussian processed labels for image scoring tasks.

3.2.2 Windows for Ordinal Classification

Moving Window Regression (MWR) [109] utilizes five neurons and a local window to estimate facial age. The model introduces the concept of relative rank (ρ rank), a novel representation scheme for input and reference instances. This relative rank is iteratively estimated by selecting two reference instances to create a search window and determining the ρ rank within that window. Essentially, MWR employs two overlapping windows with reference centers to mitigate the influence of the relative rank, also known as "intrinsic rank." Additionally, a search process is employed to identify the optimal position for the centers, reducing the impact of overlapping "rank." This innovative approach has inspired our development of a fuzzy window aimed at diminishing the overlapping features of neighboring ordinal classes.

3.2.3 Fuzzy Scoring for Ordinal Classification

Before using fuzzy logic to disjoint the characteristic adhesion between two neighbor categories, an OR-CNN is typically designed to be used for age estimation [115]. There is an expectation layer that takes the predicted distribution and label set as input and emits its expectation:

$$\tilde{y} = \sum_{k=0}^{K-1} P_k l_k, \tag{3.1}$$

where P_k denotes the prediction probability that the input image belongs to label l_k. Given an input image, the expectation regression module minimizes the error between the expected value $\tilde{y}$ and ground truth y_{true}. We use the below loss as the error measurement:

$$Loss_{err} = |\tilde{y} - y_{true}|, \tag{3.2}$$

where $|\cdot|$ denotes absolute value. Note that this module does not introduce any new parameters. OR-CNN adopts a general image classification framework that maximizes the probability of the ground-truth class during training. However, because each class is naturally influenced by its neighbors (in Figure 3.2, we can see that the 20$\sim$39 age group has a feature overlap with the 40$\sim$59 age group), the training would become unstable.

When addressing the ordinal regression problem using fuzzy logic, a notable strategy for achieving outstanding performance involves extracting a set of fuzzy rules from an example set and employing it as the foundational model with the assistance of a genetic algorithm, as demonstrated by Gamez et al. [116]. Additionally, an evolutionary fuzzy systems algorithm, leveraging monotonicity indexes, has been applied for tasks related to ordinal classification and regression [117]. Despite these advances, there remains a lack of a unified approach capable of addressing a wide range of ordinal image classification problems. This is due to the prevalent tendency among researchers to develop specific methods or systems tailored to individual problems.

Inspired by the DEX, a fuzzy scoring method has been implemented to mitigate the influence of tails and shared features within each class. This approach aims to weaken

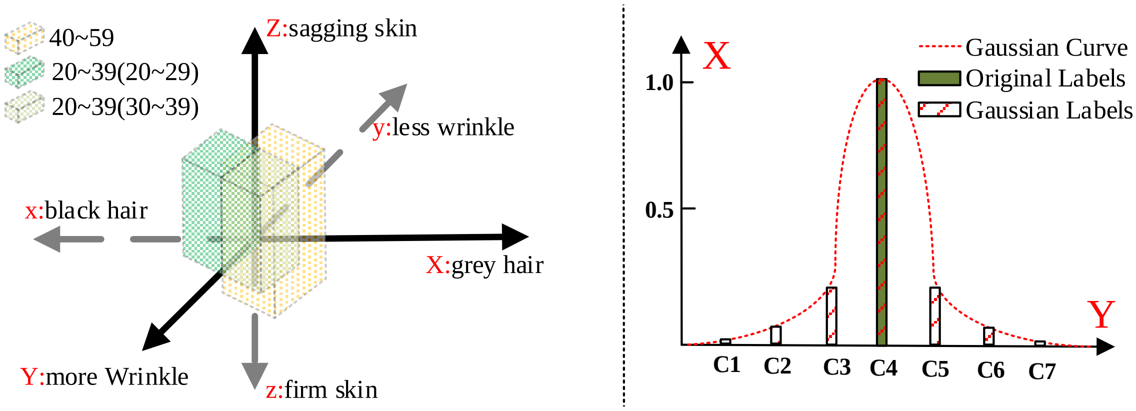


Figure 3.2: The left panel presents an example that shows overlapping features between two neighbor groups. The right panel shows the one-hot labels and the Gaussian processed labels.

feature overlaps during training steps, as presented by Kang et al. [2]. Our previous research introduced a fuzzy window that softly adjusts shared features to their optimal positions by balancing their distance from the center. For instance, in the conditions depicted in Figure 1.3a, we set the length of the fuzzy window to 3, with an ascending (or descending) trend of the high score (or low score) set at 1. Conversely, under the conditions illustrated in Figure 1.3c, to mitigate the impact of overlapping features, we adjusted the fuzzy window length to 5, with an ascending or descending trend of respective high or low scores set at 2. This configuration facilitated the forward adjustment of the output value modified by the fuzzy window towards the global average position, optimizing the redistributed probabilities.

$$\tilde{P}(x_i|y_i = i) = \frac{|i - \tilde{V}_o|}{b - a} \times \sum_{j=i-a}^{i\ +\ b} \frac{e^{-E(y_j,x_j)}}{\sum_{y_1}^{y_K} e^{-E(y_j,x_j)}}, \tag{3.3}$$

where b is the upper bound of the fuzzy window, and a is the lower bound. $\tilde{P}$ is the probability after using fuzzy windows. $E(y_j, x_j)$ is the expectation that x_j is predicted as y_j. $\tilde{V}_o$ is used to reduce the conglutination between two either neighbors or remote classes. This was calculated with:

$$\tilde{V}_o(x_j|y_j = j) = \frac{j \times P(x_j|y_j)}{\sum_{j=i-a}^{i\ +\ b} P(x_j|y_j)}. \tag{3.4}$$

3.2.4 Soft Labels and Gaussian Processes

Hard Labels: Typically represented as a one-hot vector, hard labels, such as the encoding $h_i = [0; 1; 0]$, indicate that x_i is assigned to the second class ($y_i = 2$). However, this conventional approach poses challenges in classifying ordinal images due to ambiguity and unclear boundaries. Ambiguous images complicate the determination of the appropriate class, creating an artificial gap that rigidly defines borders. This inherent drawback can impede the network's adaptability [118], [119].

Soft Labels: In contrast, soft labels represent categories using probability vectors. For instance, the encoding $h_i = [0.1; 0.7; 0.2]$ signifies that $P(Y = 2|X = x_i) = 0.7$, providing a more nuanced representation. Soft labels offer additional information to training models compared to traditional hard labels, as they convey probability distributions rather than binary classifications [118], [120]. They also possess information inheritance, enhancing resistance to disturbances during inference [104], [121], [122].

Gaussian Processes: Gaussian process approaches for ordinal regression have been explored in various contexts, including support vector machines [123], deep neural networks [45], and deep learning models with Gaussian distribution labels [124]–[126]. A study on partial label machine learning utilized the Gaussian process to handle vague labeling information, assuming an unobservable latent function dependent on the Gaussian process in the feature space of each class label [46]. However, this approach may overlook manually affected and annotated ambiguous labels, as the Gaussian distribution might not consistently represent realistic labels without logical clarification.

For facial-age detection, techniques such as Regression CNN (RCNN) [113], [127]–[129], Deep Label Distribution Learning (DLDL) [47], [104], and Deep Label Distribution Learning V2 (DLDL-V2) [40] implicitly utilize learning label distribution methods, assuming a Gaussian-like data distribution. The distribution of facial ages is modeled using a Gaussian distribution, and a lookup table is generated beforehand to store multi-part integrals that explain the probability of an input image belonging to the true chronological age [130]. In [48], label distribution learning with a normal distribution variance σ was proposed using $p_\mu(y, \sigma)$ to represent the k-th ($k \in [0, 99]$) element of $p(y, \sigma)$.

$$p_\mu(y, \sigma) = 1/\sqrt{2\pi\sigma^2} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \tag{3.5}$$

where p_μ is the probability that the true age is μ years old. It represents the connection between the classes μ and y in a normal distribution view. The optimal σ in each iteration depends on the optimal model parameter θ^*:

$$\theta^*(\sigma) = argmin_\theta L_{KL}(H, y_{true}, \theta, \sigma), \tag{3.6}$$

where $L_{KL}(H, y_{true}, \theta^*, \sigma)$ denotes the train loss. H is the training input image, while y_{true} is its label. KL is the Kullback–Leibler divergence.

3.3 Our Method

For ordinal regression, a widely adopted and effective approach involves employing multiple binary classifiers to determine the ordinal category for each input, known as the K-rank approach [40], [42], [104]. However, the success of this method hinges on the consistency of the ordinal regression data [113]. In this section, we propose a straightforward and intuitive alternative that reframes ordinal regression as a conventional classification problem. Our method utilizes Gaussian processed labels to extend the shared features between two ordinal neighbors. Subsequently, we integrate these Gaussian processed labels with a fuzzy window [2] to stabilize the weights associated with shared features.

3.3.1 Normalized Gaussian Processed Labels

After we set the equivalent double wings of the fuzzy window, which means $i-a = b-i$, we get the fuzzy $window = \{win_1 = x_{i-a}, ..., win_{a\ +\ 1} = x_i, ..., win_{a\ +\ b\ +\ 1} = x_{i\ +\ b}\}$. The true label is defined as:

$$Label(x_i|y_i = i) = \begin{cases} 0 & \text{for} \quad x_i \neq i \\ 1 & \text{for} \quad x_i = i, \end{cases} \tag{3.7}$$

and then the Gaussian processed label $Label_g$ can be:

$$Label_G(x_i|y_i = i) = \begin{cases} 0 & \text{for} \quad x_i \notin window \\ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} & \text{for} \quad x_i \in window, \end{cases} \tag{3.8}$$

where x_i is predicted to be y_i, y_i is the annotated label, i is the ordinal number, μ_o is the serially ordinal number of the true label, and σ is the variance. Here, we set $\sigma = 1/\sqrt{2\pi} \approx 0.4$ so as to ensure that when $x_i = \mu$, $Label_g$, it can be 1.

In Table 3.1, we illustrate the essential utilization difference between Gaussian Processing (GP) labels and traditional original labels. We assume there are seven categories in this example. The output probabilities of these seven categories are generated using an artificial design. In Table 3.1, the traditional back-propagation error vector (*Errors* of the output layer = probabilities of the output should be $[0.19, 0.1, 0.01, -0.6, 0.18, 0.09, 0.03]$. After we use the traditional original labels, there is only one negative error resulting from the back-propagating calculation. If, however, we apply the GP labels on the back-propagation processing, the original hard label vector will switch from $[0, 0, 0, 1, 0, 0, 0]$ to the soft label vector $[0, 0.07, 0.14, 1, 0.14, 0.07, 0]$. After using GP labels, the back-propagation error vector ($Errors_G$ of the output layer = probability outputs Gaussian labels) will turn to $[0.19, 0.03, -0.13, -0.6, 0.04, 0.02, 0.03]$. The output probability of $C3$ is lower than a systematic value—here, we assume this value was generated from the

Table 3.1: The example of using Gaussian labels. There are seven categories from C_1 to C_7, a probability vector, original labels, errors w.r.t original labels, Gaussian windows (μ = 0, and σ = 0.5), Gaussian processed labels (μ = 0, and σ = 0.5), errors w.r.t Gaussian processed labels.

Category	C_1	C_2	C_3	C_4	C_5	C_6	C_7
Probability Outputs	**0.19**	**0.1**	**0.01**	**0.4**	**0.18**	**0.09**	**0.03**
Original Labels	0	0	0	1	0	0	0
Errors	0.19	0.1	0.01	-0.6	0.18	0.09	0.03
Gaussian Window	0	0.05	0.1	0.7	0.1	0.05	0
Gaussian Labels	0	0.07	0.14	1	0.14	0.07	0
Errors$_G$	0.19	0.03	-0.13	-0.6	0.04	0.02	0.03

Gaussian function and there would be two negative errors, which, in the next step, are used for back-propagation.

The ordinal vector is $Ordinal = \{1, 2, ..., n\}$, and m is the total number of ordinal categories. Because we used cross-entropy as the loss function, the back-propagation error between the output and the last layer after using the $Label_g$ was:

$$\bigtriangledown_g(L) = |P \times Ordinal - y_i|. \tag{3.9}$$

The gradient of the weight from the α_{th} neuron in the layer $L-1$ to the β_{th} neuron in the layer L after using the $Label_g$ was:

$$\bigtriangledown_g(L-1) = \begin{cases} P \times \frac{\partial E_m}{\partial W^t_{L-1}(\alpha,\beta))} & \text{for} \quad x_i \notin window \\ \bigtriangledown_g(L) \times \frac{\partial E_m}{\partial W^t_{L-1}(\alpha,\beta))} & \text{for} \quad x_i \in window, \end{cases} \tag{3.10}$$

where E_m is the expectation output of the m_{th} category, and $W^t_{L-1}(\alpha, \beta)$ is the weight matrix of the α_{th} neuron in the layer $L-1$. We find the value of $\bigtriangledown_g(L)$ cannot always stay positive, which means when $x_i \in window$, $\bigtriangledown_g(L-1)$ should be merged using the multiplication product of $\bigtriangledown_g(L)$, $sign(P - e^{-\pi(x_i-\mu)^2})$ and $\frac{\partial E_m}{\partial W^t_{L-1}(\alpha,\beta))}$.

Figure 3.3 illustrates the distinction between utilizing GP labels and original labels. Let's consider two adjacent ordinal categories, $C_{i\text{-}1}$ and C_i, along with shared quantized features denoted by the grey area. In the initial round of Gradients Decent Directions (GDDs), the original center of the shared quantized features is positioned at $C_a(0)$. If the true label is C_i, the weight updates and back-propagation of errors cause the initial location $C_a(0)$ to shift to $C_a(1)$ (refer to Figure 3.3a). In the subsequent GDDs round, if the true label is $C_{i\text{-}1}$ based on the vector direction of the pulling force, the center will further move to $C_a(2)$. After deducing the location of $C_a(2)$, the shared features' center tends to remain proximate to either C_i or $C_{i\text{-}1}$ but not precisely at the borderline.

Conversely, when utilizing GP labels, the center of shared features exhibits fluctuations around the borderline of $C_i, C_{i\text{-}1}$. With original labels, the unidirectional pulling force of back-propagation propels the shared features' center toward C_i or $C_{i\text{-}1}$ during each

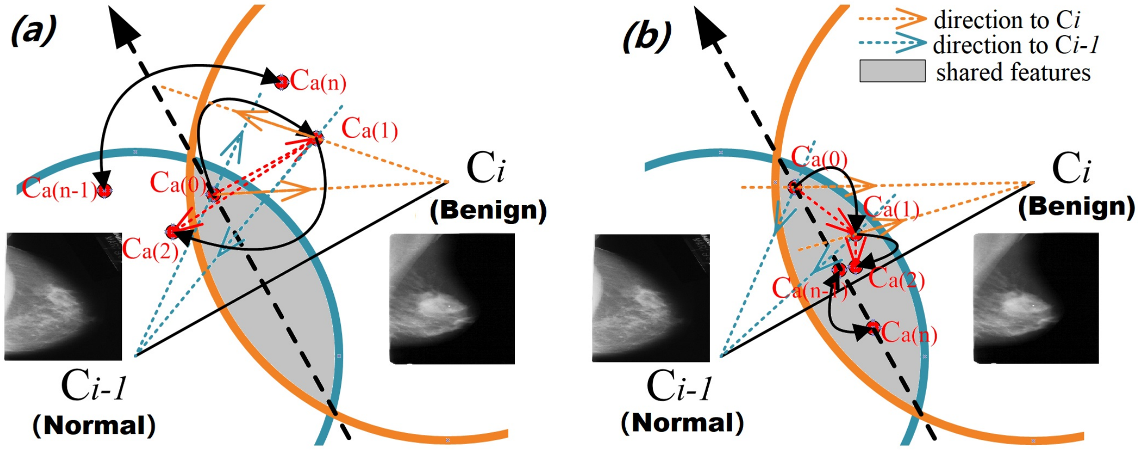


Figure 3.3: When two adjacent categories pull the center of the shared features, the resultant force decides where the center will finally stay. (**a**) When using one-hot labels, if the initial center of the shared features is $C_a(0)$, the resultant vector of the puling forces toward C_i and $C_{i\text{-}1}$ will make the center slip from $C_a(0)$ to $C_a(1)$. Finally, the center of the shared features will move close to either C_i or $C_{i\text{-}1}$. (**b**) However, if we use the Gaussian labels, the center of the shared features will finally vibrate in the middle between C_i and $C_{i\text{-}1}$.

updating step. However, when employing GP labels, the pulling force of back-propagation results from the combined influence of the C_i side and the $C_{i\text{-}1}$ side.

This part is very similar to the Fast Gradient Sign Method (FGSM) in both non-targeted and targeted adversarial attacks [131]–[133]:

$$H^{adv} = H + \epsilon \cdot sign(\bigtriangledown_H J(H, y_{non-target})) \tag{3.11}$$

$$H^{adv} = H + \epsilon \cdot sign(\bigtriangledown_H J(H, y_{target})), \tag{3.12}$$

where x is the input image, x^{adv} is the perturbed adversarial image, J is the classification loss function, $y_{non-target}$ or y_{target} is the true label for the input H and ϵ can control the steps toward to the targeted or non-targeted image. In our method, this step depends on $sign(P - e^{-\pi(x_i-\mu)^2})$, and the targeted category is the Gaussian processed neighbor of the true label.

3.3.2 Fuzzy Windows with Normalized Gaussian Processed Labels

To enhance the stability of models, it is practical to mitigate the influence of gradients by employing a lower learning rate or a smaller updating gradient. Furthermore, when contemplating the optimal global strategy involving the exclusive use of a fully connected layer, the Fuzzy Fully Connected Layer (FFCL) has been demonstrated to exert a weaker influence on the entire neural network [2]. Consequently, irrespective of the strength of the pulling force (represented by gradient matrices in each layer), the shared features' center can smoothly transition to the optimal position. This integrated approach proves particularly advantageous for classification tasks within the output OR-CNN layer.

We employ the DEX method as our foundation, wherein the true label y is quantized into distinct label groups treated as classes. In training DEX with fuzzy windows and normalized Gaussian-processed labels, we substitute the expectation module (the final output layer) with fuzzy windows of varying lengths. We utilize a Gaussian function ($\sigma = 1/\sqrt{2\pi} \approx 0.4$) to process ordinal labels and subsequently modify the loss function with a conventional cross-entropy loss. The back-propagation error between the output and the last layer, following the application of the $Label_g$, is expressed as:

$$\tilde{\nabla}_{x_i} l(x_i, y_{true}) = \begin{cases} \tilde{P} - 0 & \text{for} \quad x_i \notin window \\ \tilde{P} - Label_g & \text{for} \quad x_i \in window, \end{cases} \tag{3.13}$$

where $\tilde{P}$ is calculated from Equations (3.3) and (3.4).

In Algorithm 1, we present the pseudo-code for the fuzzy window with a normalized Gaussian processed label algorithm, addressing the ordinal regression issue. The initial step involves processing labels through a Gaussian distribution. Upon defining the length of the Gaussian window, L_{Win}, the calculation of $Label_G$ follows the Gaussian processing template outlined in Table 3.1. However, when the Gaussian window extends to the beginning or end of the entire age sequence (0 or m), any out-of-range elements (e.g., if the front side of the window $Frt < 0$ ($Frt = i - L_{hWin}$) or if the back side of the window $Bk > m$ ($Bk = i + L_{hWin}$)) should be excluded.

The second step employs fuzzy logic to mitigate the impact of overlapping features in ordinal neighbor classes. The value $\tilde{P}_i$ can be computed using Equation (3.3), and $\tilde{V}_o$ is determined through Equation (3.4). During the inference phase of the fuzzy window, an anticipated value, obtained by summing the product of two elements—the position of the binary classifier and the prediction probability of that specific classifier—is utilized for the final estimation.

3.4 Experiments

In this section, we introduce one medical image dataset CBIS-DDSM and four facial-age datasets (IMDB-WIKI, FG-NET, MORPH-2, CACD, and Depression EEG). In the following, there are three experimental ablation results. The first shows the performance on the selection of the hyperparameter L_{Win}. The second ablation study presents the

Algorithm 1 Fuzzy Windows with Normalized Gaussian Processed Labels

Input: Gaussian Processed Labels with Windows

The true labels y_i=$\{2, 10, ..., 99\}$, label matrix $Label(y_i) = \{Y_1, Y_2, ..., Y_n\}$. Set binary matrix $Label(y_i) = zeros(m, n)$, m is the length of categories, n is the number of samples, and set the length of the Gaussian window as L_{Win}. Because L_{Win} is an odd number which is greater than 1, the half length of the Gaussian window $L_{hWin} = 0.5 \times (L_{Win} - 1)$. The output probability of the model is P_i, and the ordinal vector of the fuzzy window is $Ordinal_i = \{y_i - L_{hWin} + 1, ..., y_i, ..., y_i + L_{hWin} - 1\}$. Initialize probability $\tilde{P}_i = P_i$;

Output: Four Variables

Initialize $Label_G(y_i)$=$\{Y_{G1}, Y_{G2}, ..., Y_{Gn}\}$; $\tilde{P}_i$; $\tilde{V}_o$;
$Error_G(y_i)$=$\{Err_{G1}, Err_{G2}, ..., Err_{Gn}\}$.

for $j = 1; j \leq n$ **do**
- compute $Frt = i - L_{hWin}$;
- compute $Bk = i + L_{hWin}$;
- **if** ***Normalized Gaussian Processed Labels*** **then**
  - Compute $G_W = e^{-\pi k^2}$, and $k = \{-(L_{hWin} - 1), -(L_{hWin} - 2), ..., 0, ..., L_{hWin} - 2, L_{hWin} - 1\}$. Initialize $Label_g(y_i) = Label(y_i)$
  - **if** $Frt < 0$ **then**
    - replace Y_{Gi} with $[G_W(|Frt| + 1), ..., G_W(L_{Win}), 0, ..., 0]$;
  - **else if** $Bk > m$ **then**
    - replace Y_{Gi} with $[0, ..., 0, G_W(1), G_W(L_{Win} - |Bk|)]$;
  - **else**
    - replace Y_{Gi} with $[0, ..., 0, G_W(1), ..., G_W(L_{Win}), 0, ..., 0]$, where the index of $Y_i(G_W(1))$ is $i - L_{hWin} + 1$;
- **else if** ***Fuzzy Windows*** **then**
  - Compute $\tilde{V}_o$ by Equation (3.4); Compute $\tilde{P}_i$ by Equation (3.3);
- **else**
  - Continue;
- $Error_G(y_i) = \tilde{V}_o$ - y_i ;

return $Label_G$, $\tilde{V}_o$, $\tilde{P}_i$, $Error_G$;

Table 3.2: Sample distribution of CBIS-DDSM dataset based on BI-RADS assessment.

Scores (BI-RADS)	**0**	**1**	**2**	**3**	**4**	**5**
Training Set (Mass + Calcification)	192 (129 + 63)	1	559 (77 + 482)	368 (279 + 89)	1286 (533 + 753)	458 (299 + 159)
Testing Set (Mass + Calcification)	46 (33 + 13)	2	85 (14 + 71)	109 (85 + 24)	347 (169 + 178)	115 (75 + 40)

performance of FW-GPL in processing a designed fragmentary ordinal dataset. The last one demonstrates comparison results with SOTA methods on three facial-age datasets.

3.4.1 Datasets

In this study, there are one medical image dataset and four different facial-age estimation datasets (one for pretraining).

Ordinal Medical Dataset

Table 3.2 shows the size of one ordinal medical dataset and its corresponding splits for training and testing.

CBIS-DDSM Dataset. The CBIS-DDSM comprises a substantial collection of digitized film mammography images, encompassing 3572 images corresponding to 2689 patient cases. The dataset classifies cases based on the BI-RADS, utilizing assessments ranging from 0 to 5. These assessments include BI-RADS score 0 for incomplete cases, BI-RADS score 1 for negative cases, BI-RADS score 2 for benign cases, BI-RADS score 3 for probably benign cases, BI-RADS score 4 for suspicious abnormal cases, and BI-RADS score 5 for highly suspicious malignant cases. Access to the CBIS-DDSM dataset is available at `https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM`.

Depression EEG Dataset. The EEG data associated with depression severity includes labels determined through the Structured Clinical Interview for DSM-IV, the 17-item HAMD, and scores provided by two professional clinical psychologists. The dataset comprises 52 healthy undergraduate dextromanual students (with a gender distribution of $6:4$ males to females and a mean age of 20.4 ± 9.7) and 48 depressed patients (with a gender distribution of $6:4$ males to females and a mean age of 34.3 ± 12.1). Following established criteria [38], three groups were defined: healthy controls (non-depressed, scores 0-7), depressed with low scores (scores 8-23), and depressed with high scores (scores $\geq$24). The dataset is accessible at `https://github.com/ChengKang520/Classifying-and-Scoring-MDD-BCI`.

Facial-Age Estimation Datasets

Table 3.3 shows the size of each dataset, and the corresponding splits for training and testing.

Table 3.3: Facial-age datasets used to evaluate the proposed FW-GPL.

Datasets Name	Train	Test	Val	Total	Label Range
IMDB-WIKI	260,282	⊗	⊗	523,051	0–100
FG-NET	990	12	⊗	1002	0–69
MORPH 2	4380	1095	⊗	5475	16–70
CACD	145,275	10571	7600	163,446	⊗

IMDB-WIKI. For the IMDB-WIKI dataset (IMDB-WIKI can be downloaded from `http://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/`), the authors crawled images of celebrities from IMDB (`www.imdb.com`) and Wikipedia (`https://en.wikipedia.org/`).

FG-NET. The Face and Gesture Recognition Research Network (FG-NET) [107] aging database consists of 1002 color and grey-scale images, which were taken in a totally uncontrolled environment. On average, there are 12 images for each of the 82 subjects, whose age ranges from 0 to 69 (FG-NET is available at `https://yanweifu.github.io/FG_NET_data/`).

MORPH-2. The Craniofacial Longitudinal Morphological Face Database (MORPH) [106] is the largest publicly available longitudinal face database containing more than fifty thousand mug shots (You can find MORPH-2 from `https://www.faceaginggroup.com/morph/`).

CACD. The Cross-Age Celebrity Dataset (CACD) [108] collected from the Internet contains 163,446 images from 2000 celebrities. This dataset splits into three parts, 1800 celebrities are used for training, 80 for validation, and 120 for testing (The link of CACD is `http://bcsiriuschen.github.io/CARC/`).

3.4.2 Evaluation Metrics

For model evaluation and comparison [134], we computed the Mean Absolute Error (MAE) [135] and RMSE [136], on the test set after the last training epoch:

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|\tilde{y} - y| \tag{3.14}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}|\tilde{y} - y|^2} \tag{3.15}$$

where $\tilde{y}$ is the output value of OR-CNNs, y is the real facial age label, and N is the total number of test samples.

3.4.3 Experiment Settings

Following DEX [112], Soft Stagewise Regression (SSR) [43], Mean Variance (MV) Loss [41], and Compact yet efficient Cascade Context-based Age Estimation (C3AE) [42], the model can be first pre-trained on the IMDB-WIKI dataset. This method can be embedded into any CNN ordinal classification model. We respectively set the length of the fuzzy window L_{Win} as 10 for facial-age detection and 3 for breast cancer detection. We used the Adam optimizer in all the experiments, and similarly to SSR and C3AE, the initial learning rate, dropout rate, momentum, and weight decay were set to 0.002, 0.2, 0.9, and 0.0001, respectively. The learning rate was 0.001 with a decay every 10 epochs by a factor of 0.9. Compared with the SOTA methods, each model totally trained two hundred epochs with a batch size of 50. During the training steps, to avoid overfitting the overlapping features, we adjusted the training strategy according to Algorithm 2.

Algorithm 2 Training $Model$

Input: The accuracy rate Acc after validation in every epoch, threshold ratio is $Tre_{ratio} = 0.8$, and $epoch_N$ is the number of total epochs. Before using Algorithm 1, $Acc_{average}$ is the average accuracy;

Output: $Model$.

```
for j = 1; j ≤ epoch_N do
    if Current Acc_j > Tre_ratio × Acc_average then
        Train Model under Algorithm 1 with FW-GPL;
    else
        Train Model under Algorithm 1 without FW-GPL;
return Model;
```

3.4.4 Hardware and Software

All loss functions and neural network models were implemented in MATLAB2019b and PyTorch 1.7 and trained on four Tesla V100 graphics cards (The source code is available at `https://github.com/ChengKang520/FW-with-GPL-for-Ordinal-Regression`.

3.5 Results and Analysis

So as to compare with the SOTA results, we respectively summarize the comparison result of CBIS-DDSM in Table 3.4 and the comparison result of facial-age detection in Table 3.5.

3.5.1 Scoring Breast Cancer Images

Table 3.4: Comparison with existing methods on DDSM in terms of ACC.

Method	CNN + FCL	CNN + FFCL	CNN + W-GPL
Geras [137] (BI-RADS: 0/1/2)	68.8%	70.1%	70.3%
Akselrod-Ballin [138] (BI-RADS: 2/(3-4-5))	60.0%	62.3%	62.4%
Kang [2] (BI-RADS: 0/(2-3)/(4-5))	72.0%	74.1%	74.2%
Kang [2] (BI-RADS: 0/1/2/3/4/5)	56.34% ±1.4%	57.40% ±1.7%	58.29% ±1.9%

As we set the hyperparameter $L_{Win} = 3$ when scoring BI-RADS, only when the categories are greater than 3 in number can our FW-GPL work well to reduce the influence of overlapping features among neighboring ordinal classes—this can also be seen in Table 3.4. We find that only when scoring the BI-RADS of six categories does FW-GPL show a weak but obvious improvement. The distance d between BI-RADS score 2 (benign) and BI-RADS score 3 (probably benign) is probably beyond the "boundary", as is the distance d between BI-RADS score 4 (suspicious abnormal) and BI-RADS score 5 (highly suspicious malignant); therefore, the classification task for BI-RADS is difficult.

3.5.2 Scoring Facial-Age Images

When scoring facial-age images, we configured the model with 10 neurons and a window length of 5. Table 3.5 provides a summary of the results obtained by comparing our model with SOTA models across three facial-age datasets. In contrast to label distribution learning methods like DLDL-V2 [40] and MV Loss [41], our FW-GPL utilizes a fixed pattern (Gaussian processed labels) to learn features that account for age distribution without requiring prior knowledge of the image data's age distribution. Compared with models employing specialized loss functions, especially FW-GPL demonstrates competitive performance against most SOTA methods, including MV Loss [41], SSR [43], and C3AE [42]. The effectiveness of FW-GPL can be attributed to the fuzzy window, which mitigates the influence of conjugation among neighboring ordinal categories. Unlike DLDL-V2 [40], MV Loss [41], and SSR [43], which consider the entire probability, or C3AE [42], which focuses solely on the two highest output probabilities.

Another advantage lies in the fact that the Gaussian processed labels eliminate the need to fit a hyperparameter σ [48] to approximate the true age probability distribution. In comparison to FW-GPL, MWR [109] employs global and local relative ordinal regressors (ρ regressors) to predict ρ ranks within both the entire and specific rank ranges. Additionally, MWR refines an initial search window iteratively, moving it by selecting two reference instances, and ultimately estimates the ρ rank within the window.

Table 3.5: In terms of MAEs, our approach is compared with different SOTA methods. (* indicates the model was pre-trained on the IMDB-WIKI dataset.)

Type	Method	MORPH 2	FG-NET	CACD	Paras
Bulky	**DEX** [112]	3.25	4.63	-	138M
	DEX * [112]	2.68	3.09	6.52	138M
	MV [41]	2.41	4.10	-	138M
	MV * [41]	2.16	2.68	-	138M
	DLDL-v2 [40]	1.969	-	-	138M
	FP-Age [139]	2.04	5.60	5.60	138M
	FP-Age * [139]	***1.90***	4.68	***4.33***	138M
	DRF [44]	2.80	3.47	5.63	-
	PML [140]	2.31	***2.16***	-	-
	JREAE [51]	2.71	3.390	4.596	-
	MWR [109]	*2.13*	-	5.68	-
	FW-GPL [Ours]	2.71	4.27	-	138M
	FW-GPL * [Ours]	2.24	2.73	6.10	138M
Compact	**ORCNN** [41]	3.27	6.44	-	479.7K
	MRCNN [41]	3.42	-	-	479.7K
	SSR [43]	3.16	-	-	40.9K
	C3AE [42]	2.78	4.09	-	39.7K
	C3AE * [42]	2.75	2.95	-	39.7K
	AVDL * [48]	2.37	2.32	-	11M
	MWR [109]	***2.00***	***2.23***	-	-
	FW-GPL[Ours]	2.72	3.71	-	40.9K

3.5.3 Scoring Depressive Severity using EEGs

Table 3.6: Comparison with existing methods on scoring depressive severities with EEGs.

References	Subjects	Cross validation	Method + Feature	RMSE	MAE
Images (Scenario)					
Kosuke Yoshida et al (2017)[89]	MDD = 58, HC = 65	leave-one-out	PLS + sMRI	9.56	-
Benson et al (2012)[103]	MDD = 30, HC = 0	leave-one-out	RVR + MRI	2.50	-
EEGs (Scenario)					
Hashempour et al (2022)[141]	MDD + HC = 119	10-fold	CNN-TCN + EEGs (64 electrodes) + Eyes-Open State	2.37±1.3	1.73±0.27
Kang et al (2023)[6]	MDD = 48, HC = 52	10-fold	ResNet + EEGs (beta bands 16 electrodes) + N-back Paradigm	2.80±1.6	2.01±0.32
The proposed method	MDD = 48, HC = 52	10-fold	ResNet + EEGs (beta bands + 16 electrodes) + N-back Paradigm + FW-GPL ($win = 10$)	2.41±1.5	1.87±0.34

After we set the hyperparameter $L_{Win} = 10$, our FW-GPL works well to score depressive severity, as it achieves the lower RMSE and MAE comparing to SOTA results. With the use of EEGs, Hashempour et al [141] also applied CNNs under Temporal-Convolutional Neural Network (TCN) to score depressive severity. But they used 64 electrodes, not 16 electrods to detect depressive severity. In Table 3.6, the application of EEGs and CNN models presents the best performance.

3.6 Ablation and Discussion

Based on the facial-age image classification, we used the ordinal IMDB-WIKI data to do the ablation analysis. The ablation study was conducted in three parts: (1) to analyze the influence of the number of neurons, (2) to analyze the influence of the length of window L_{Win}, and (3) to figure out how this model could process incomplete ordinal data.

3.6.1 Ablation Study I (Influence of the Number of Neurons)

We used the classical pre-trained DEX model as the base. In Tables 3.7 and 3.8, we see that when the neuron number N is 10 or 5, the DEX model can get the best performance. This finding echoes prior research showing that when the number of neurons in the output layer is 10 or 5, DEX-family age detection models can achieve better performance [109]. In other words, a smaller N has a better error tolerance.

Table 3.7: Test performance of the FW-GPL method, with the $L_{Win} = 10$ (set length of output neurons N as [100, 50, 20, 10, 5]).

Method	**DEX**	**DEX with FW-GPL**			
$\boldsymbol{N}$	100	50	20	10	5
RMSE	12.46	13.36	12.65	12.60	12.80
MAE	8.94	8.67	8.79	8.62	8.59

Table 3.8: Test performance of the DEX method (set length of output neurons N as [100, 50, 20, 10, 5]).

Method	**DEX**	**DEX without FW-GPL**			
$\boldsymbol{N}$	100	50	20	10	5
RMSE	13.57	13.38	12.86	12.67	12.71
MAE	8.96	8.83	8.77	8.64	8.74

3.6.2 Ablation Study II (Influence of the Length of the Window L_{Win})

We employed two types of output layers (corresponding to $N = 100$ and $N = 10$) to assess the performance of FW-GPL across different values of L_{Win}. The results are consolidated in Tables 3.9 and 3.10. Since the length of the half window, L_{hWin}, must exceed the number of neurons in the output layer, we observed that a broader window ($L_{Win} = 50$ or $L_{Win} = 100$) yields superior performance when there are 100 neurons in the output layer. Similarly, this trend is consistent when the number of neurons in the output layer is 10. The rationale behind this observation is that a wider window can accommodate ample information for accurately estimating facial age. However, when $N \leq L_{hWin}$, no discernible improvement is observed. This underscores the importance of selecting an appropriate window length, as it directly influences the performance enhancement of the FW-GPL model.

Table 3.9: Test performance of FW-GPL on testing data sets (length of output neurons set as 100).

Method	DEX with FW-GPL				
L_{Win}	5	10	20	50	100
RMSE	15.17	15.10	14.58	13.65	13.68
MAE	10.18	10.11	9.78	9.69	9.71

Table 3.10: Test performance of FW-GPL on testing data sets (length of output neurons set as 10).

Method	DEX with FW-GPL		
L_{Win}	5	10	20
RMSE	12.91	12.60	12.60
MAE	8.78	8.62	8.62

3.6.3 Ablation Study III (Incomplete Ordinal Image Data)

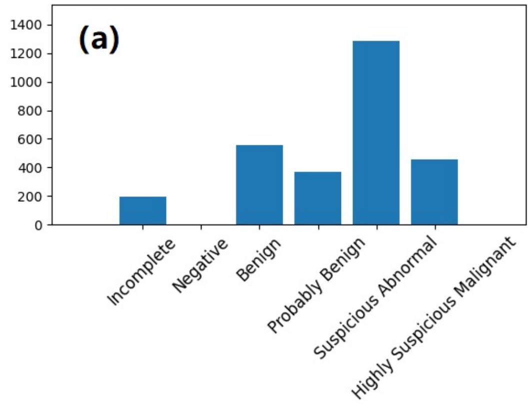
We manually removed some age segments of the IMDB-WIKI to train the model and test it in the complete ordinal text data, as shown in Figure 3.4. In Table 3.11, we can see that when the number of neurons is 100, the most proper window is 20. In Table 3.12, when we set the length of the window as 10, the lowest MAE appears when the number of neurons is 5. Consequently, there is no obvious difference between the incomplete (this section) and complete (Ablation Study II) ordinal image data, and the result can only be affected by the number of neurons N and the length of the window L_{Win}.

Table 3.11: Test performance between the DEX and the FW-GPL with fragmentary IMDB-WIKI dataset. (length of output neurons set as 100).

Method	DEX	FW-GPL			
L_{Win}	0	5	10	20	50
RMSE	12.46	14.73	14.30	13.64	12.72
MAE	8.94	9.43	9.13	**8.78**	8.81

3.6.4 Advantage and Limitation

In addressing the challenge of ordinal image classification head-on, our approach aims to mitigate the impact of overlapping features. The window's length plays a pivotal role in controlling the defuzzification of ordinal neighbor categories. While JREAE [51] leveraged two covariance matrices to capture correlations in both input facial features and output



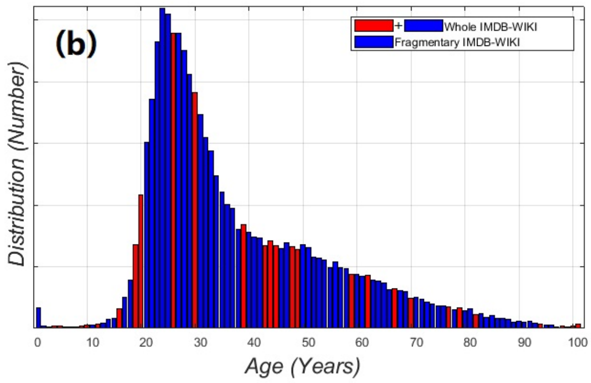


Figure 3.4: This figure shows the condition that the BI-RADS or the facial-age dataset is not consecutive. (**a**) The class distribution of CBIS-DDSM. (**b**) The age distribution of the IMDB-WIKI. The blue bars are the fragmentary IMDB-WIKI, whereas the red bars are manually removed.

Table 3.12: Test performance between the DEX and the FW-GPL with fragmentary IMDB-WIKI dataset. (length of the window set as 10).

Method	**DEX**		**FW-GPL**		
N	101	50	20	10	5
RMSE	12.46	13.36	12.65	12.60	12.80
MAE	8.94	8.67	8.79	9.08	**8.59**

age labels, certain methods within this family (e.g., DRF [44] and AVDL [48]) should first consider the age distribution inherent in the dataset. This consideration becomes crucial because fitting the distribution of the facial age dataset may introduce an inevitable deviation from the true age distribution. To circumvent this issue, our method employs a Gaussian distribution within the window to approximate the relationship between input facial features and output age labels. As evidenced in Table 3.5, our approach outperforms other LDBL methods, showcasing the advantages of employing label distribution-based learning techniques.

However, a notable drawback lies in our reliance on a simple fuzzy logic window to tackle the challenges posed by ordinal image classification tasks. Unlike MWR [109], which dynamically adjusts the distance between the real age and the center of the moving window, our method confines the window's center using naive fuzzy logic to modify the facial age distribution within the window. This approach tends to overlook the influence of remote yet highly relevant features lying beyond the window's scope. Despite attempts to utilize longer windows, our method struggles to address this limitation effectively.

3.7 Conclusions

In this paper, we propose a novel approach for ordinal image scoring, termed Fuzzy Window with Gaussian Processed Label Learning (FW-GPL). FW-GPL introduces a method to mitigate the impact of overlapping features between two ordinal neighbors, resulting in superior performance compared to other methods across multiple age estimation datasets and a medical dataset with ambiguous annotations. Our experiments also demonstrate that FW-GPL effectively handles discontinuous ordinal regression by adjusting the window length.

The innovative use of fuzzy logic and a Gaussian process strategy to guide ordinal image classification is promising, and we plan to explore additional applications. Several directions for future work are identified: (1) We intend to apply this method to various ordinal medical tasks, such as scoring the severity of depression and grading spinal cord injuries, in our forthcoming research. (2) Despite not achieving the best State-of-the-Art (SOTA) result, we aim to enhance performance by integrating FW-GPL into other SOTA models. (3) To optimize computational efficiency, we will fine-tune pre-trained models incorporating FW-GPL to minimize computing costs.

Chapter 4

Brain Networks of Maintenance, Inhibition and Disinhibition During Working Memory

WM, a crucial cognitive function responsible for information maintenance, serves as an indicator of brain function. Activities associated with memory sustention, inhibition, and disinhibition are integral to understanding the fundamental neurocognitive architecture. Despite proposed brain models attempting to elucidate the entire WM process, conclusive evidence and detailed descriptions, particularly regarding the regions and structures involved in maintenance, inhibition, and disinhibition, remain sparse. In our study, we utilized phase lock coherence and general partial directed coherence to establish connections among four adaptively fitted EEG sources. Additionally, we employed previously published models to characterize the brain circuits associated with maintenance, inhibition, and disinhibition. Conducted with forty-five mental health undergraduates using a classical visual n-back paradigm, our experiment revealed that the bilateral PFC primarily focused on cognitive components such as rehearsal before recognition for object classification, inhibition to maintain positive memory and activities, and disinhibition to timulate subsequent interactions in the brain. Furthermore, our findings indicated that the right PFC occasionally assisted the left PFC in executing high-capacity WM tasks. In contrast, posterior regions, specifically the PPC, were observed to be engaged in attention arousing and maintenance. These two key observations suggest that: **a)** the recurrent maintenance circuit plays a vital role in executing positive cognitive components; **b)** the instantly monitoring inhibition temporarily pauses the sustained function to conserve energy; and **c)** the arrival of disinhibition stimulates the next step in the brain, prompting the selection of new subjects or a focus on novel subjects.

4.1 Introduction

WM is defined as the capacity to guide behaviors [142] and has been linked to the control of attention [143] and academic performance [144]. Although there is no consensus on the neurocognitive architecture of WM, its fundamental conceptualization involves the short-term maintenance of information [145]. Numerous traditional WM paradigms, characterized by lower capacity, have been employed in various clinical populations to assess the performance of individuals with mental impairments, including schizophrenia, stroke, traumatic brain injury, and Attention Deficit-Hyperactivity Disorder (ADHD). A critical yet unmet clinical need exists for non-invasive measures to evaluate WM activity and guide psychological interventions.

This study undertakes a multifaceted approach. Firstly, we (i) evaluate behavioral performance through the implementation of n-back paradigms. Subsequently, (ii) we analyze brain networks associated with WM using phase-lock coherence and directional coherence after adapting a 64-channel EEG and generating four sources to simulate cerebral internal communications. Additionally, (iii) we propose a "neurocognitive architecture" of working memory based on region-to-region connections, revealing pathways of memory maintenance and lateral inhibition during WM. This study sheds light on the processes of WM and its corresponding brain regions through coherence and provides a non-invasive assessment of functional networks during WM tasks in the healthy population.

The proposed neurocognitive architecture of working memory [145] encompasses **1)** the selective attention process, **2)** object information recognition and maintenance, **3)** rehearsal process, **4)** update and attention sustentation, and **5)** inhibition [145], [146]. This concept, building on existing ideas, elucidates the WM procedure by integrating various processing descriptions and emphasizing the concepts of memory maintenance and lateral inhibition [145]. Visual cortex, prefrontal cortex (PFC, primarily comprising the posterior superior frontal gyrus and middle frontal gyrus), posterior parietal cortex (PPC, predominantly located in the intraparietal and superior parietal cortex), and inferior temporal cortex are integral regions in visual WM paradigms [142], [147], [148]. This chapter delineates the processes of WM and its associated regions through brain coherence and introduces a non-invasive assessment of functional networks during WM tasks in the healthy population. Based on different neurocognitive stages, four major procedures during WM tasks are elucidated in this chapter.

1. We assessed the behavioral performance after subjects implementing n-back paradigms,
2. we examined brain networks of WM by phase-lock coherence and directional coherence after the 64 channels EEG adaptively fitted and four sources generated to simulate cerebral internal communications,

3. We proposed our "neurocognitive architecture" of WM based on region-to-region connections, and found the pathways of memory maintenance and lateral inhibition during WM.

4.2 Related Work

4.2.1 Pathway for Attention Arousal and Executive Function

It has been proposed that the Prefrontal Cortex (PFC) plays a crucial role in resilient information maintenance during WM tasks. Numerous meta-analyses have consistently shown that the left PFC, particularly the ventral aspect, is closely associated with verbal WM tasks, while activation of the right PFC, particularly the dorsal aspect, is consistently observed in spatial WM tasks [149]–[152]. Lesion studies affirm these associations, revealing electrophysiological activities that demonstrate neural connections in the PFC of monkeys [153], [154]. Additionally, fNIRS has been employed to assess WM load by monitoring blood activities in the PFC [155], further establishing the importance of PFC in normal WM. Alongside the PFC, the Parietal Posterior Cortex (PPC) is strongly implicated in WM tasks [156]. Spatial WM tasks, which typically activate the right hemisphere, engage the bilateral parietal cortex [149], [151]. Subsequently, both fMRI and Positron Emission Tomography (PET) studies have demonstrated that the PFC can select content represented in posterior regions [152]. Nevertheless, some studies have proposed that the superior parietal cortex may be associated with executive function and selective attention control [157], [158]. Furthermore, investigations into the integrity of white matter pathways have elucidated connections among the PFC, parietal cortex, and temporal cortex during WM tasks [148], [159].

4.2.2 Pathway for Coding and Decoding

Working memory necessitates the encoding and subsequent selection of relevant content amidst distractors [160]. The interplay between the PFC and PPC has been validated as a source of top-down signals that insulate stimulus-coding networks [161], [162]. The adaptive coding observed in PFC showcases its fundamental capability to classify learning tasks [163], [164]. Notably, the phenomenon of population coding within PFC neurons has been identified as contributing to the transition between various representational states, particularly in the context of a delayed paired associates task [165].

Sources analysis has revealed the initial dynamic visual encoding occurring in posterior brain regions, alongside the encoding of selection rules in the prefrontal cortex. These encoding and decoding components, as identified through sources analysis, play a crucial role in the maintenance of memory content [166]. Furthermore, multivariate decoding and source analyses have provided insights into the reliance on prefrontal and parieto-occipital persistent oscillatory neural activity for the selection of memory content [166].

4.2.3 Pathway for Sustained Brain Activity

Maintenance and sustention in brain might consist of memory storing, goals and tasks keeping, and attention sustaining. Stronger synaptic connectivity were thought to be associated with the brain network of sustained higher activity [145]. Particularly, fronto-parietal activity was examined to be relative to components of task-general processing, such as maintaining goals and task sets [167].

4.2.4 Pathway for Lateral Inhibition

In tasks involving WM, it is imperative to consider multiple factors for inhibition, particularly when the WM system reaches its capacity. Inhibition becomes essential to prevent the decay of persistent activity [142]. A dynamic model employing a winner-take-all mechanism has been proposed to elucidate robust lateral inhibition among memory representations, highlighting that inhibition typically leaves only the winning representation active [145]. To integrate both new and old information in WM tasks, cognitive inhibition (the ability to inhibit irrelevant information and selectively attend to goal-relevant information) and response inhibition (the ability to inhibit a prepotent response) are crucial components [168]. The attribution of these distinct regions to storage or executive components and their communication during the implementation of relative WM tasks remained unclear [151]. Regions such as the superior parietal cortex, posterior part of the superior frontal area, and the middle frontal area are implicated in completing visuospatial WM tasks and tasks requiring the aforementioned components [161], [162], [169]. Despite recent identification of active regions by researchers, few studies have explored the architecture of inhibition and sustention among these areas, particularly the structure of loops explaining WM. To validate the "human neurocognitive architecture" of working memory, we utilized EEG sources and their connections to construct a communicational model based on these cognitive components. Various dynamic and statistical algorithms, such as the directionality of neural information flow [170]–[172] and PLC relying on time lag [8], [173], [174], have been developed to measure the transmission of neural signals. Approaches like Partial Directed Coherence (PDC) [175] and GPDC [176] have been proposed for analyzing brain networks based on EEG studies. Both PLC and GPDC have demonstrated significant utility in structural systems [177] and in real organisms, including applications in Parkinson's [172] and Alzheimer's [178] disease patients,

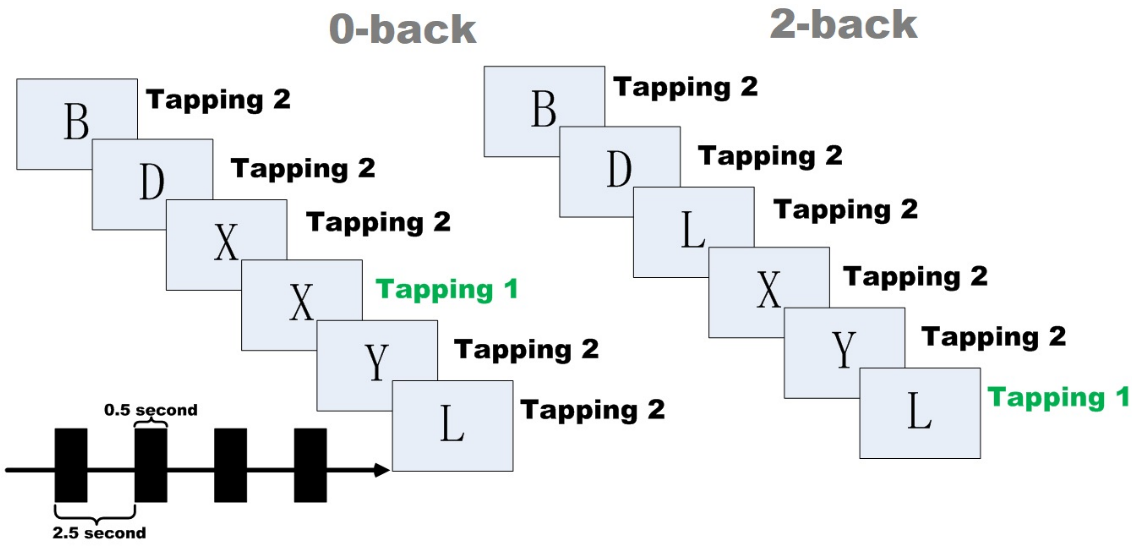


Figure 4.1: The experimental procedures with their timelines. Subjects should respond to stimuli by pressing the number key 1 with index finger for match (target stimulus) and pressing the number key 2 with middle finger for mismatch (nontarget stimulus).

depression patients [7], [8], hippocampal-prefrontal activation in monkeys [177], and kainic acid-anesthetized rats [179].

4.3 Methods

4.3.1 Participants

Forty five healthy undergraduate dextromanual students (6:4 male to female, and mean 20.4 years) were called for visual n-back paradigms. This study was approved by the local institutional ethics review board. A written informed consent was obtained from each subject before the experiment. These subjects have no medication before, and no personal or family history of psychiatric or neurological disease.

4.3.2 Experimental Procedures

We utilized E-Prime 5.0 to design the n-back experiment for our study. The letter variant version of the n-back tasks, encompassing 0-back as a baseline and 2-back as the working memory load, was employed. Participants were instructed to observe and respond to stimuli on the screen by pressing the index finger button for a match (target stimulus) and the middle finger button for a mismatch (nontarget stimulus). Specifically, during 0-back tasks, subjects were tasked with identifying a pre-specified letter 'X,' while in 2-back tasks, they were required to recognize a letter that matched the one presented two trials back. The presented letters were randomly selected from English consonants (as illustrated in Figure 4.1).

The entire experiment was segmented into three parts, each comprising two 0-back tasks and two 2-back tasks, with the task order randomly arranged. The duration of each 0-back or 2-back task was set at 75 seconds, featuring a pseudorandom sequence of 30 consonants (10 targets and 20 nontargets). Letters were presented for 0.5 seconds, followed by a 2-second disappearance to allow subjects time for reaction. A 45-second break separated each part. Subjects were instructed to respond as swiftly and accurately as possible. For behavioral performance analysis, reaction time and response accuracy were recorded, with incorrect responses excluded from the EEG analysis. Prior to the formal experiment, subjects underwent practice sessions, repeating tasks until they were confident in their understanding of every detail.

4.3.3 EEG Recording

The EEG data were recorded with the BrainAmp amplifier (Brain Products, Munich, Germany) and Braincap electrode cap (EASYCAP, Herrsching, Germany). According to the international 10–20 system, all 64 Ag/AgCl channels were referenced during recording to electrode (FCz) with a forehead ground (AFz). To remove eye movements, vertical and horizontal Electrooculogram (EOG)s were recorded from two additional channels located at the right side of the right eye and below the left eye. Electrode impedance was maintained below 5kΩ throughout the experiment. No filter was used during recording (Sampling rate: 1000 Hz).

4.3.4 Data Analysis

After a band-pass filter at 0.16-30 Hz (24dB/Octave), artifact rejection and baseline correction, EEG data analysis was divided into data preprocessing, source modeling, phase lock coherence and general partial directed coherence. The aim of data preprocessing is to acquire standard trials for each subject. After artifact rejection and removal of trials that subjects responded incorrectly, an average of 53 trials for 0-back tasks and an average of 49 trials for 2-back tasks were kept among all subjects, which were used to construct source model and coherence analysis. The PLC provides stable connections within specific durations, whereas GPDC can display the direction of each connection and the detail in different time periods.

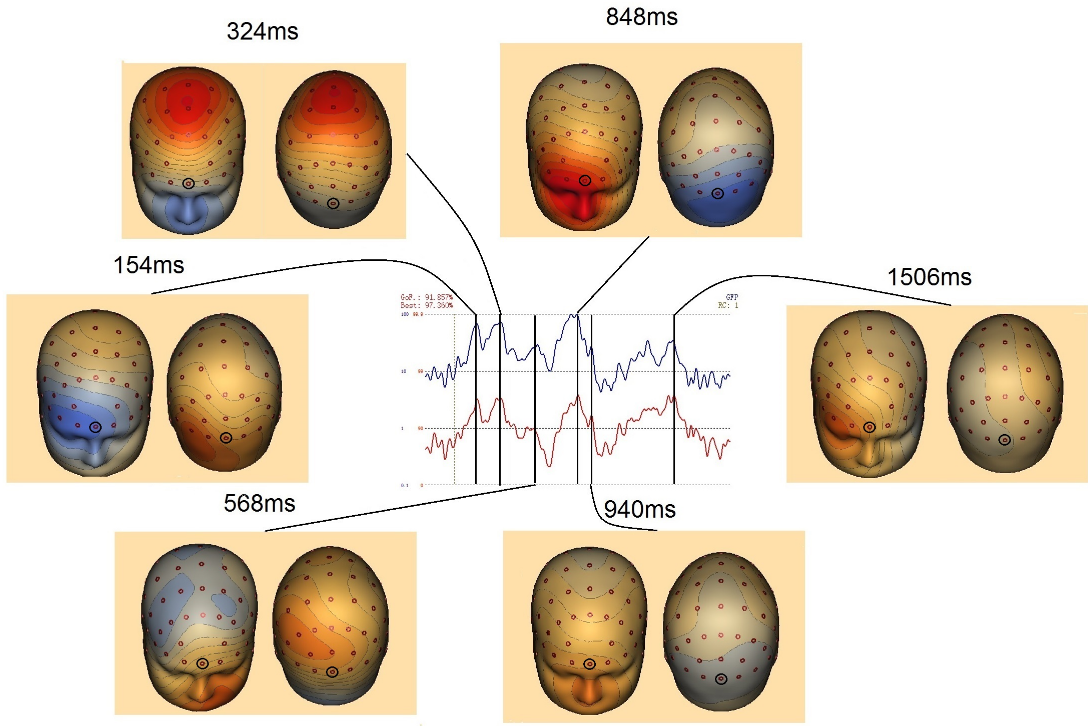


Figure 4.2: Illustration of brain altered scalp voltage maps of 2-back condition minus 0-back condition in the front and back hemispheres during different periods. The circled electrode sites are Fz and Oz. The GFP (the sum of squares of all channels, normalized to 100%) shown in the central is displayed in a logarithmic scale..

Data Preprocessing and Single-Trial Source Waveform Extraction

The averaged Evoked Related Potential (ERP) waveforms for individual subjects were computed under both 0-back and 2-back conditions. Subsequently, the difference wave between these two conditions was calculated for each subject. By averaging the differences across all subjects in each corresponding channel, the collective representation of EEG was generated, illustrating scalp topography performance (see Figure 4.2). This representation was then transformed into source waveforms (Figure 4.3). Building upon fMRI findings that highlighted activations in the bilateral superior/inferior parietal lobules and bilateral inferior frontal gyri in the 2-back vs. 0-back contrast [180], a discrete model with four sources was developed to extract source waveforms. The decision to construct a Regional Source (RS) model instead of a dipole model was based on the capability of a single RS, composed of three mutually orthogonal dipoles, to accurately represent activity in multiple gray matter patches with different orientations in a specific brain region [181], [182]. Additionally, an inter-hemispheric symmetry constraint for the coordinates of the RS pair was imposed during source modeling. By inputting the difference waveforms between 2-back and 0-back conditions into Brain Electrical Source Analysis

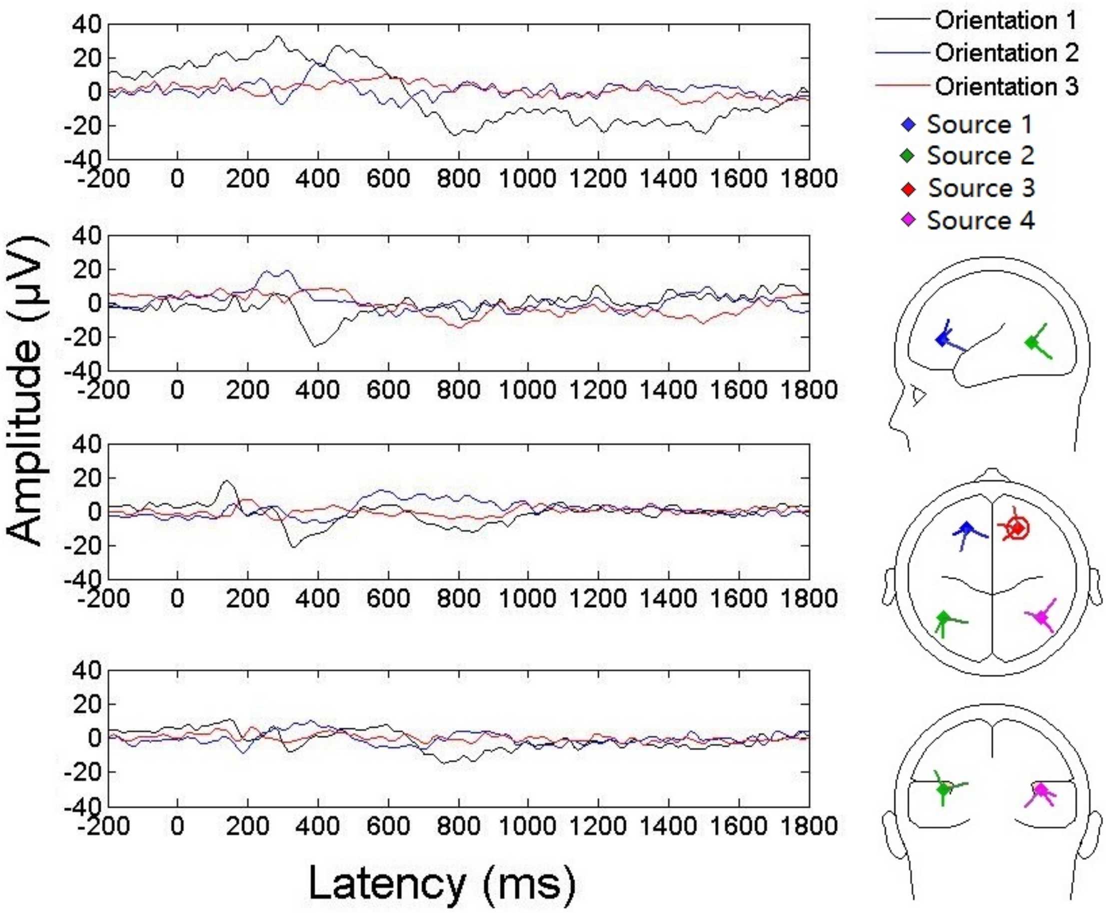


Figure 4.3: RSs and their corresponding time courses of the group average EEGs. The three directional time courses of RSs are displayed in the left panel, meanwhile, their locations and orientations of RSs are presented by using three head views in the right panel. Orientation 1 is the primary orientation of each RS. Four sources were generated to simulate the difference waves between 2-back tasks and 0-back tasks.

software (BESA 6.0) and aligning them with the best correspondence between recorded and estimated scalp distribution, we calculated the source configuration within a realistic head model. The primary orientation of each RS was set to match the direction of the maximum dipole moment (main current flow) of the averaged difference ERP waves (Figure 4.3). Subsequently, the resulting RS model was applied to EEG data during 0-back and 2-back tasks, extracting single-trial source waveforms for each subject. Furthermore, the activity obtained in the primary orientation of each RS (orientation 1 in Figure 4.3) was utilized for coherence analysis.

PLC Analysis

For phase locked coherence analysis, a complex Morlet's wavelet (For computation in EEGLAB, the number of wave cycles was set to 0.5s, and the lowest frequency time window to 0.5 sec) was used to calculate the time-frequency domain:

$$\omega_{trial,i}(f,t) = \left(\frac{1}{\sqrt{\pi}\delta_t} \exp\left(-t^2/2\delta_t^2\right) \exp(j2\pi ft)\right. \tag{4.1}$$

where $\omega_{trial,i}(f,t)$ is the product of a sinusoidal wave at frequency f and time t during trial i, with a Gaussian function with a standard deviation δ_t. We defined the strength of phase synchrony as phase lock value ($PLV_{,m}$) between RS l and m with the following equation [7], [177]:

$$PLV_{l,m}(f,t) = \left|\frac{1}{n}\sum_{\text{trail}=1}^{n} \exp\left(i\left[\omega_{\text{trail },l}(f,t) - \omega_{\text{trail },m}(f,t)\right]\right)\right| \tag{4.2}$$

where n is the number of available trials. $PLV_{l,m}(f,t)$ is computed by 1Hz steps from 1Hz to 30Hz. The set of $PLV_{l,m}(f,t)$ is termed Phase Lock Value (PLV) below. To identify the task-dependent modulation of the PLV, a typical two-sample t test was applied to test the significant difference of PLV between 0-back and 2-back in terms of latency and frequency domains. An one-sample t-test was perfor6med on the acquired t-values from the two-sample t-tests to determine the task-dependent modulation of the PLV across the subjects [8], [173], [174]. Moreover, 1000 times of bootstrap re-sample We also measured the directional coherence through phase lag, and the calculation of mean phase lag between each two sources is:

$$\varphi_{(l,m)}(f,t) = \text{ angle } \left\{\frac{1}{n}\sum_{\text{trail }=1}^{n} \exp\left(i\left[\omega_{\text{trail },l}(f,t) - \omega_{\text{trail },m}(f,t)\right]\right)\right\} \tag{4.3}$$

A circular bootstrap test was used to test whether the distribution of phase lags across all sources was significantly different from zero.

GPDC Analysis

As consistent phase lags much smaller than a full oscillatory cycle are suggestive of directional influences, they are in principle ambiguous because of the cyclic nature of the signals. We measured the GPDC [176] value among these four generated sources to measure the directed connections. It can measure causality by predicting one signal from past values of another signal in terms of the degree (GPDC value). This method based on a type of P-order Multivariate Autoregressive (MVAR) model:

$$X(t) = \sum_{p=1}^{P} A_p(n)X(t-p) + e(t) \tag{4.4}$$

where A_p is the autoregressive coefficient matrix with the size of 4×4 and p is time lag,

P is the maximum number of lags (model order), $X(t)$ is the concatenated matrix of four source signals at time t, and $e(t)$ is the residual error vector. The MVAR model order P can be calculated by evaluating and where M is the number of time series, P is the optimal model order, N is the time point and σ is the covariance matrix. The MVAR coefficients can be obtained by two different ways [170]: **1)** the mean coefficients of all single-trial MVAR coefficients, and **2)** the MVAR coefficients of the data concatenated from all single-trial source waveforms. We selected the second way to calculate the MVAR coefficients, and set each sliding time window as 2000 ms with 50 ms step between successive windows during different trails and tasks conditions. According to our previous study [170], we employed Kalman smoother method [183] to figure out the optimal estimator for MVAR coefficients, which only can rely on previous measurements and inevitable time lag.

The fitted MVAR parameters were then transformed from the time domain into the frequency domain:

$$\Lambda_{l,m}(f,t) = I - \sum_{p=1}^{P} A_p(t) e^{-j2\pi fp/F_s} \tag{4.5}$$

where I is the $p \times p$ identity matrix, with the sampling rate F_s in terms of $(l \to m)_{th}$ entry, and $\Lambda_{l,m}(f,t)$ were evaluated from $1 \sim 30$Hz at every 1 Hz step. The value of GPDC then indicating the directional connections among these four sources is calculated as:

$$GPDC_{l \to m}(f,t) = \frac{|\Lambda_{l,m}(f,t)|}{\sqrt{\sum_{m=1}^{M} |\Lambda_{l,m}(f,t)|^2}}, l = 1, \ldots, M, \quad m = 1, \ldots, M, \tag{4.6}$$

where $\Lambda_{l,m}(f,t)$ is the variance of the prediction error for order P. After the calculation of GPDC, the two sample t-test was used again to identify the significant time-frequency domain between baseline (0-back) and 2-back. Although 1000 times of bootstrap re-sample method was employed again and scattered significant areas were drawn with gray band (95% confidence interval level), we still sorted out the significant area through 5×5 median filter, and pick out some obvious time-frequency domains. The bootstrap method can detect the time-frequency regions, where the GPDC values in 2-back tasks are significantly different compared to those values in 0-back tasks. To address the problem of multiple comparisons, the significance level (p value) was corrected using a False Discovery Rate (FDR) procedure.

4.4 Study Results

4.4.1 Behavioral Result

We recorded the subject's behavioral performance during tasks implementing. In Table I, both response accuracy ($p < 0.001$) and reaction times ($p < 0.001$) of these two tasks did significantly differ between groups.

4.4.2 Scalp Topography Performance

After group-averaging the waveforms, we employed a contrast between the 2-back and the 0-back (baseline) conditions. Four distinct peaks are evident in Figure 4.2. The initial peaks manifest at 158 ms and 324 ms, coinciding with a transition in scalp topographic activity from the left temporo-occipital lobe to the centroparietal lobe. Notably, prefrontal hyperactivity is observed between 844 ms and 1328 ms, indicating a shift of activated areas towards major frontal regions. Further details reveal a reduction in frontal potential from 848 ms to 1328 ms, with concurrent activation of the prefrontal, frontal, and temporal lobes.

4.4.3 Band-Specific Synchrony Reflects

We conducted a thorough examination and validation of communication through phase-locking synchrony among the four sources depicted in Figure 4.4. Prior to 700 ms, as illustrated in Figure 4.4a and 4.4b, the connection between S2 and S3 exhibited highly synchronized coherence, with the left PPC lagging behind the right PFC (mean rel. $phase$ =-17.20, $p < 0.001$, $r = 0.943$, bootstrap test versus zero phase lag; Figure 4.4a middle panel). This synchronization was prominent in the late theta and early alpha bands ($6 \sim 11$ Hz). Simultaneously, a strongly concentrated phase coherence was observed for the posterior connection (mean rel. $phase$ =-4.21, $p < 0.001$, $r = 0.875$, bootstrap test versus zero phase lag; Figure 4.4a right panel) in the late beta-band ($28 \sim 29$ Hz). In the phase-locked activities after 700 ms, as depicted in Figure 4.4c and 4.4d, notable connections include the front connection between S1 and S2 (mean rel. $phase$ =-17.91, $p < 0.001$, $r = 0.833$, bootstrap test versus zero phase lag; Figure 4.4c right upper panel) during late alpha and early beta bands ($11 \sim 16$ Hz), the left lateral connection between S1 and S3 (mean rel. $phase$ = 11.08, $p < 0.001$, $r = 0.946$, bootstrap test versus zero phase lag; Figure 4.4c left under panel) during the middle beta band ($17 \sim 22$ Hz), and the right lateral connection between S2 and S4 (mean rel. $phase$ = 14.89, $p < 0.001$, $r = 0.790$, bootstrap test versus zero phase lag; Figure 4.4c right under panel) during early and middle beta bands ($14 \sim 19$ Hz, and $21 \sim 26$ Hz).

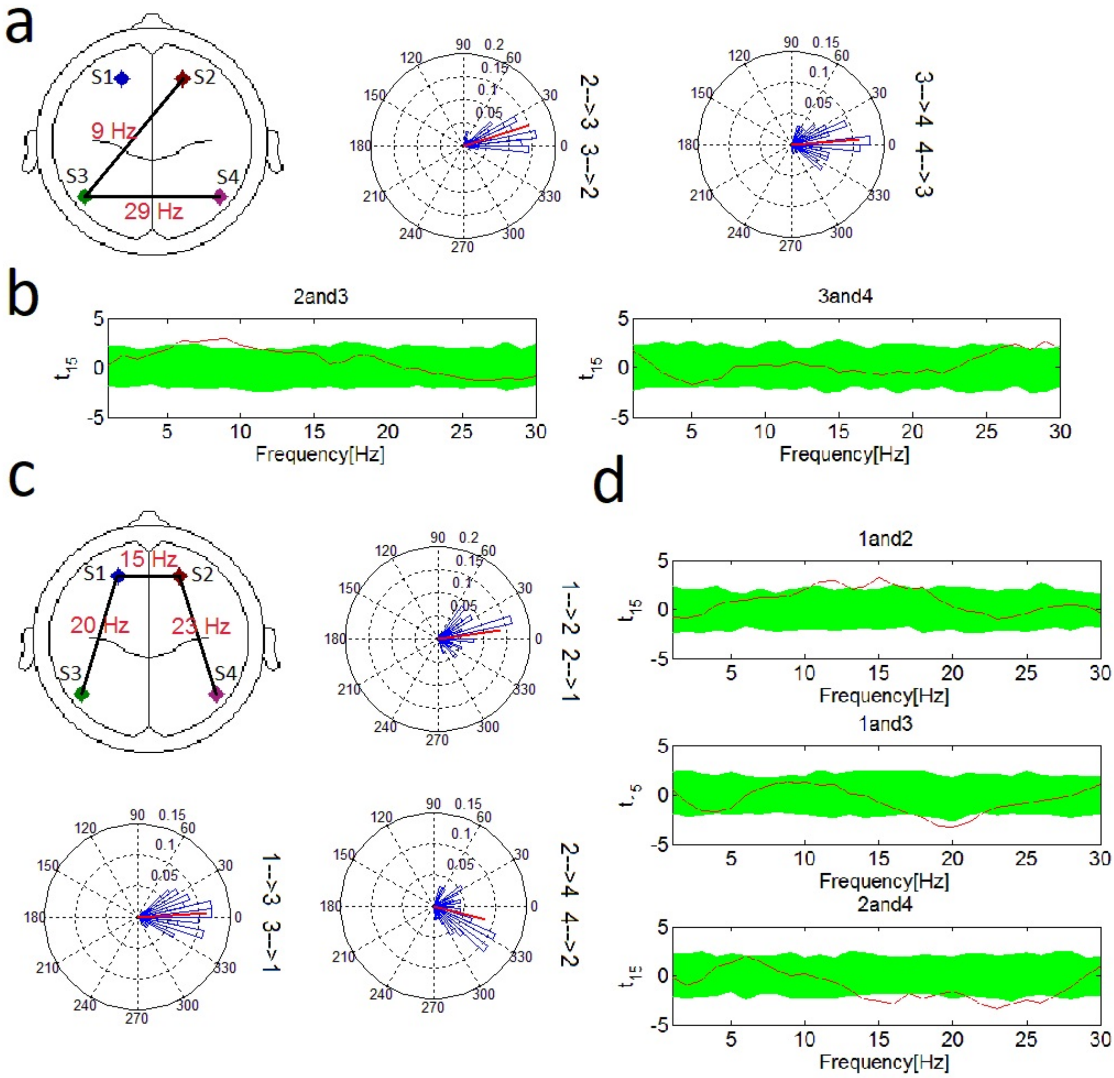


Figure 4.4: The phase locked connections among these four sources from 0 ms to 700 ms (**a**, **b**), and from 700 ms to 1600 ms (**c**, **d**). **(a)** Left panel shows the connections under some particular frequencies, and right panel presents circular statistic angles with their distribution, circular histograms also illustrate the mean angles of the phase differences between specific two sources presented (red line). (**b**, **d**) t statistical value for the difference in PLV between 2-back and 0-back tasks for RS pairs across subjects. For example, in the pair of S1 and S3, the PLV in the 18 Hz $\sim$ 21 Hz beta band was higher in 2-back tasks with a peak at 20 Hz, green band is the t value of one-sample t test with 95% confidence interval constructed using the bootstrap method, and red line is the t value. **(c)** the connections under some particular frequencies, and circular statistic angles with their distribution.

4.4.4 Band-Specific Directionality Reflects

This directed coherence differed in the direction of putative causal influence appears within whole frequency bands from theta band to beta band. Figure 4.5a illustrates time-frequency regions exhibiting significantly increased GPDC. The significant time-frequency domain respectively was presented in Figure 4.5a, and the directed connections according to diverse neurocognitive processes were drawn in Figure 4.5b. When 2-back tasks were compared with 0-back tasks, connection E (150∼300 ms), connection D (550∼700 ms) were detected before responses. After responses, by contrast, Figure 4.5b presents connection A and F (700∼900 ms), connection C (900∼1100 ms), connection H, B and G (1300∼1600 ms). There is no significance between 0-back and 2-back tasks after 1600 ms, and the duration is 2000 ms, therefore, the last procedure is neglected in our study.

4.4.5 The Neurocognitive Architecture With Component Processes of WM

Drawing insights from recent research [142], [145] utilizing fMRI and electrophysiological methods, Figure 4.6 depicts the involvement of specific cognitive components. Notably, selective attention manifests during the P300 duration [7], [184], verbal rehearsal is evident [185], sustained activities occur [186], and retrieve/readout processes are engaged [145], [187]. Additionally, pattern recognition [186], memory update and storage [188], and lateral inhibition [142], [145] contribute to the cognitive landscape.Preceding responses, posterior connections play a pivotal role in arousing selective attention. Bilateral prefrontal regions host rehearsal and retrieve/readout processes, while sustained attention and pattern recognition unfold between crossed right-prefrontal and left-parietal regions, occurring within the 500 ms $\sim$ 700 ms timeframe after an initial silent period of approximately 250 ms. Post-responses, sustained attention and lateral inhibition unfold in the anteroposterior right hemisphere. Simultaneously, updating and memory encoding processes take place in bilateral prefrontal regions during the 700 ms $\sim$ 900 ms interval. From 900 ms to 1100 ms, a repetition of cognitive and memory components serves to maintain brain activity in visual WM tasks. Between 1100 ms and 1600 ms, sustained attention monitors targeted objects, and lateral inhibition mitigates the risk of failure.Finally, we propose a novel neurocognitive architecture for WM processing in Figure 4.7, addressing the gaps in current WM explanations. This architecture comprises directed arrows and loops, providing a comprehensive depiction of WM processes.

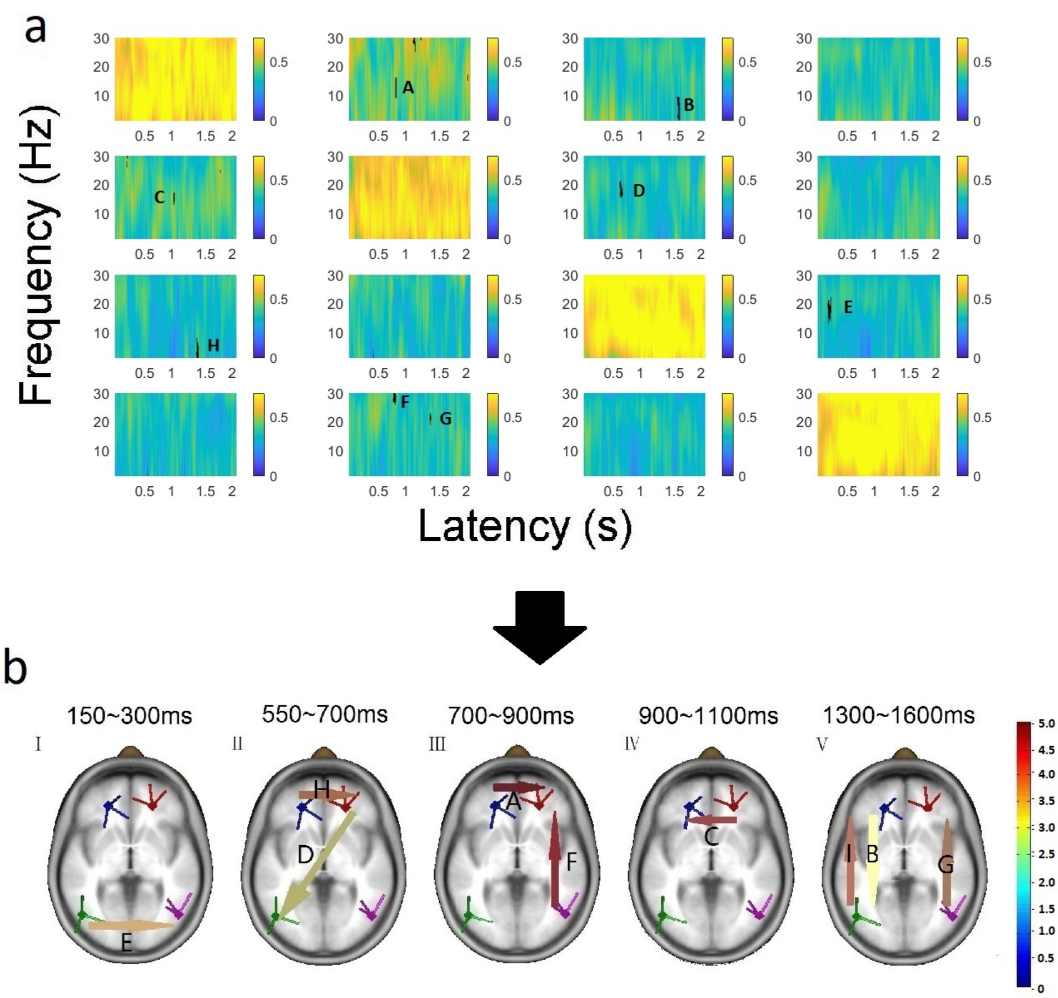


Figure 4.5: The directed connections based on the time-varying GPDC. **(a)** Time-frequency representations of the time-varying GPDC under 2-back tasks accompanying with significant grey blocks through two-sample t test between 0-back and 2-back. The bar presents the value of GPDC. **(b)** According to directed connections in different latencies, directed color arrows shows the information flow and their strength thereof. In the earlylatency interval (I: about 150-300 ms E, and II: about 550∼700 D), the cortical contacts mainly include S3→S4 E from 10 to 25 Hz and S2→S3 D, and both of these two indicate the transmission of trigger information. In the late-latency interval (III: about 700∼900 ms, IV: about 900∼1100 ms, and V: about 1300∼1600 ms), the cortical information was transmitted by S1→S2 A between 12 Hz and 17 Hz, S2→S1 C between 12 Hz and 22 Hz, S4→S2 F between 25 Hz and 30 Hz, S1→S3 B between 1 Hz and 14 Hz, S3→S1 H between 1 Hz and 6 Hz, and S4→S2 G between 17 Hz and 23 Hz.

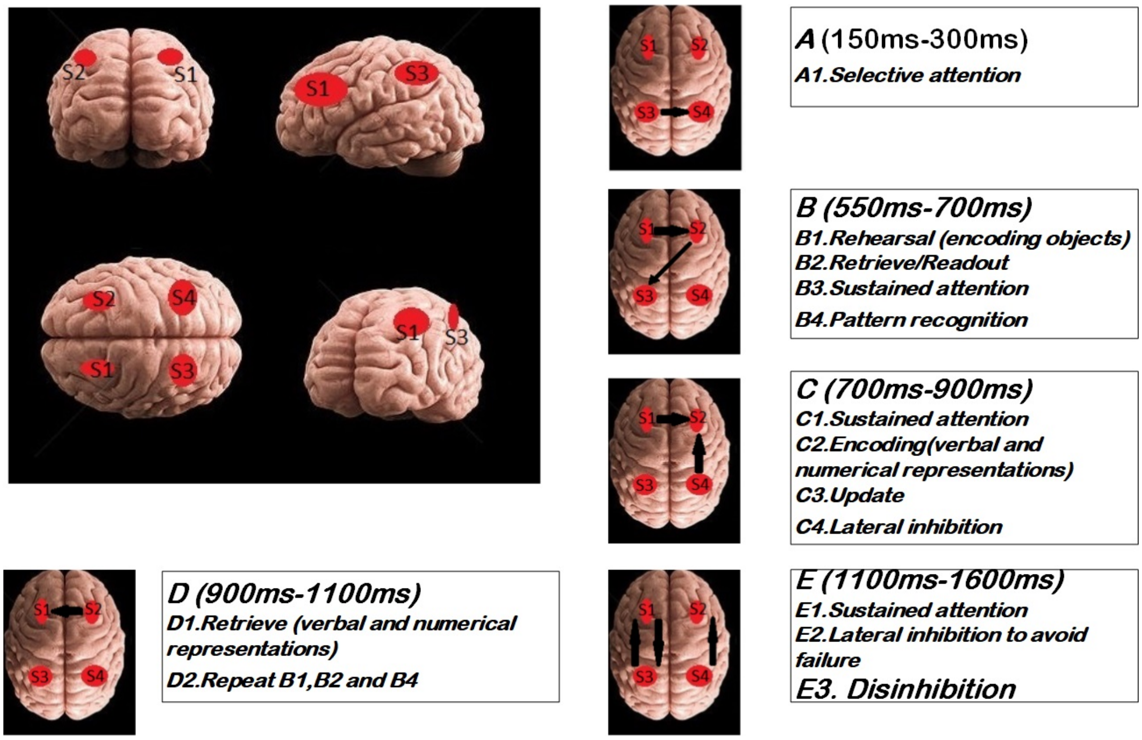


Figure 4.6: Schematic explanation of representations to brain networks during WM tasks. Left upper panel is the location illustration of four fitted sources. **A**$\sim$**E** present components relative to WM in terms of some specific neurocognitive processes. **A.** During this duration, selective attention is activated by the trigger of capitals shown on the screen, and this induced the attention mechanism in PPC cortex. **B.** Executive and cognitive functions between right PFC cortex and left PPC region, appear after selective attention being implemented to process numerical and verbal information. **C.** The PFC and right hemisphere connections indicate the update of information flow for memory storing, and lateral inhibition to avoid the failure of memory representation. **D.** Persistence of information under WM tasks happens in PFC cortex. **E.** The last process for the recall of sustained attention, lateral inhibition to avoid the failure of attention and memory processing, as well as disinhibition.

4.5 Discussion

In the traditional visual n-back paradigm employed in this study, two coherence methods were employed to construct the brain network during WM tasks. These methods adaptively identified four sources primarily situated in the bilateral PFC and PPC, regions associated with functions related to working memory. Specifically, under the condition of 2-back minus 0-back, PLV depicted undirected connections, while GPDC illustrated directional connections. Remarkably, both coherence methods revealed a similar network structure. Based on these findings, we proposed a comprehensive model for the working memory process. The model incorporates unique directional cognitive and executive connections, along with two regular cognitive and memory maintenance cycles. Preceding responses, the targeted stimulus initially triggered selective attention in the parietal regions and was subsequently encoded in visual cortex areas. The beta posterior connections in Figure 4.4a and the broad beta directional causality in Figure 4.5b-I indicated the arousal of attention, coinciding with the fixation of the target on the screen. Contrary to initial speculation, beta oscillations, particularly in terms of selective attentional control, appeared to govern attention and top-down processing [189]. Building on Eriksson's suggestion of a core circuit involving fronto-parietal cortical regions sustaining attention and supporting rehearsal [145], our analysis fused the primary alpha coherence in Figure 4.4a with the beta directional connection in Figure 4.4b (D). This fusion suggested a rehearsal simulation occurring between the right PFC and left posterior parietal cortex (PPC). Despite the different frequency bands employed in these two methods during rehearsal, the brain network appeared capable of generating an early-stage simulation rooted in internal reasoning. Consequently, the trigger initiated attention arousal, and the visual cortex encoded the target while transmitting the representational information to the attention and rehearsal networks.

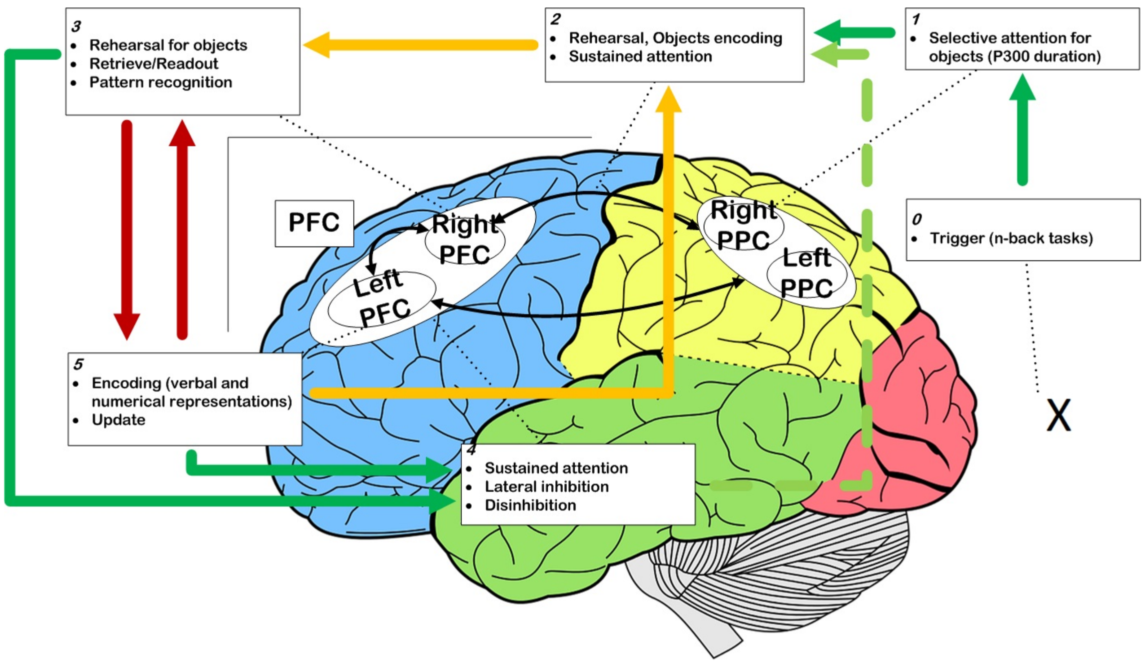


Figure 4.7: Summary of our proposed neurocognitive architecture for WM. **X** is the visual n-back trigger. Before responses, there are attention arouse link (**0-1-2**), activity maintenance loop (**2-3-2**). Attention arouse always accompanies with object encoding to recognize the type of objects. After response, brain maintenance loop mainly consists of activity loop (**2-3-5-2**) and major memory loop (**3-5-3**), and inhibition or disinhibition loop (**2-3-4-2**, **2-3-5-4-2**), as the core joint is (**4**). Inhibition networks are crucial to guarantee the accuracy of information and activity in brain, meanwhile, disinhibition circuits are important to restart the following brain activities. Therefore, activity loop and major memory loop need inhibition component.

4.5.1 The Maintenance Loop During WM

Frontal regions have been identified as crucial during delay periods, wherein activity profiles unfold over time [190]. Previous research suggests that sustained frontal brain responses in working memory tasks are linked to selection mechanisms rather than the encoding of memory content [191]. Additionally, meta-analyses indicate that the left PFC, particularly the ventral region, is more associated with WM tasks, while the right PFC is more involved in spatial WM tasks [149]. In our n-back paradigm experiments, the directional connection D in Figure 4.4b-II may elucidate the flow of information from the right PFC to the left PPC. We propose that, in our experiments, the right PFC may have functioned as a buffer storing information, subsequently read out to facilitate information selection and comparison during retrieval.

For the maintenance of information or activities in the brain, short-term maintenance requires sustenance functions to support sustained brain activity during WM tasks. The period from 300 ms to 550 ms, characterized by a relative absence of significant activities, suggests a relatively stable activation of the brain during WM execution. Drawing on cognitive findings [192], [193], we speculate that this "silent period" from 300 ms to 550 ms may resemble a P300 component. Although our experiment did not delve into the mechanism of this silent period during WM tasks, we hypothesize that this silence may reflect preparatory contributions to subsequent high-level WM processes. Some researchers propose that after such a "silent period," sustained top-down influences transform information representations during WM to guide final decisions [166].

The red loop between bilateral PFCs in Figure 4.7 signifies the maintenance of memory information post-responses. This maintenance loop between left and right PFCs may activate a self-reminder function, enhancing memory or ensuring necessary brain activity. Recent fMRI analyses involving older and younger adults performing WM tasks revealed weaker BOLD signal increases in the DLPFC among older individuals during relative maintenance [194]. This study emphasized the need for both manipulation and maintenance, where attention is directed to transition to the next position in the alphabet sequence while maintaining the result in working memory. The yellow loop denotes the enhancement of short-term memory through repeated rewriting, comparison, and correction. Additionally, the representation of a capital sequence during WM requires cognitive management to rank at least three capitals in the brain. While rehearsals can keep information in the brain, sustained performance necessitates cyclic repetition, possibly relying on auxiliary cortex regions, with the right PFC potentially acting as a compensatory region for memory storage scheduling, temporary storage preparation, and sustained attention.

4.5.2 The Inhibition Loop During WM

WM, previously believed to rely on the interplay of recurrent excitation among pyramidal neurons for sustaining persistent activity through a delay (silent period), as well as lateral inhibition to moderate interneurons and diminish the impact of external distractions [195]–[197], involves the ability to curtail the activation of functionally similar neurons within its local environment, particularly through lateral inhibition [198]. In the context of filtering distractor input, lateral inhibition becomes hyperpolarized when background noise threatens to influence neuronal firing [199]. Recent insights underscore the role of lateral inhibition within posterior areas and the prefrontal cortex for resource-limited descending [198]. Effective performance in tasks involving memory and positive brain activities necessitates the control of extraneous information or functions. The inhibition of unnecessary activities is crucial not only for the filtering of distractors but also for the arousal and maintenance of attention. The observed lateral inhibition following the red and yellow loops aligns with a recent theory positing that attention and working memory capacities are constrained by flexible cortical connections that process overlapping inhibitory surrounds, potentially leading to internal competition for completing WM tasks [200].

Instances of forgetting may occur due to an overall decrease in firing frequency resulting from inadequate inhibition, insufficient to maintain recurrent activity. Our experimental evidence weakly supports the notion of prefrontal lateral inhibition during WM, particularly after the delay period, aiming to preserve memory representations and prevent the shifting of brain attention and task-focusing functions. Nevertheless, we speculate that lateral inhibition may more frequently induce cognitive stagnation, causing a deadlock in information maintenance, brain responses, and the activation of subsequent positive or negative neuron circuits.

To initiate subsequent brain activities, disinhibition becomes crucial, especially as the delay period increases. This mechanism is supported by a study on schizophrenia during spatial WM experiments [199]. Cortical disinhibition's effect on WM may serve as the switch to activate the next loop of maintenance or brain activity, as the absence of disinhibition could lead to disorder during the implementation of a specific WM component.

4.5.3 Conclusion And Future Directions

We have elucidated the phase lock and directional connections among four adaptively fitted sources to conduct noninvasive coherence analysis. Building upon the established architecture of WM, we have proposed a detailed network model encompassing maintenance and inhibition. This model underscores the significance of disinhibition, particularly highlighting its partial functions in bilateral PPC regions. The analysis of this model yields several key findings: **(i)** Bilateral PFC and PPC are crucial in WM tasks, contributing to attention, rehearsal, recognition, inhibition, and disinhibition; **(ii)** The right PFC acts as a facilitator for the left PFC, enhancing the high-capacity implementation of WM tasks; and **(iii)** Following inhibition for maintenance in the brain (in one loop to enhance memory or sustain positive activities), disinhibition unlocks the inhibitory function, activating subsequent brain functions. In our future work, we will concentrate on identifying abnormal connections and addressing the unbalanced WM observed in depressed patients.

Chapter 5

InA: Inhibition Adaption On Pre-trained Language Models

Fine-tuning pre-trained language models (LMs) may not always be the most practical approach for downstream tasks. While adaptation fine-tuning methods have shown promising results, a clearer explanation of their mechanisms and further inhibition of the transmission of information is needed. To address this, we propose an Inhibition Adaptation (InA) fine-tuning method that aims to reduce the number of added tunable weights and appropriately reweight knowledge derived from pre-trained LMs. The InA method involves **(1)** inserting a small trainable vector into each Transformer attention architecture and **(2)** setting a threshold to directly eliminate irrelevant knowledge. This approach draws inspiration from the shunting inhibition, which allows the inhibition of specific neurons to gate other functional neurons. With the inhibition mechanism, InA achieves competitive or even superior performance compared to other fine-tuning methods on $BERT-large$, $RoBERTa-large$, and $DeBERTa-large$ for text classification and question-answering tasks.

5.1 Introduction

Fine-tuning, the process of updating the parameters of pre-trained LMs, has proven to be an effective approach for various downstream NLP tasks. However, classical fine-tuning methods suffer from the issue of redundant parameters in fully pre-trained models, which can lead to inefficiencies when adapting to new downstream tasks. To tackle this problem, prior studies have attempted to adapt only specific vectors or learn additional parameters while keeping most of the pre-trained parameters fixed. This allows for better operational efficiency by loading task-specific parameters associated with the pre-trained models before deployment. Low rank adaption (LoRA) ([52](has successfully achieved this goal and addressed the inference latency problem, which helps extend model depth or reduce the usable sequence length of models ([53]–[55]) to find a balance between efficiency and quality. The challenges in fine-tuning pre-trained LMs for NLU downstream tasks lie in reducing the number of tuned weights and appropriately approximating the update of pre-trained weights derived from the LMs ([52], [53], [55], [56]). Properly selecting knowledge from pre-trained LMs is crucial to address these challenges. The question arises as to why we cannot directly inhibit "redundant" knowledge during fine-tuning while retaining relevant information.

Drawing inspiration from the efficiency demonstrated in neural networks by [201], and the concept of low 'intrinsic rank' in weight changes during model adaptation proposed by LoRA ([52]), we propose our approach called Inhibition Adaptation (InA). Our hypothesis is that by partially inhibiting the intrinsic rank, we can eliminate the influence of irrelevant 'intrinsic parts' in the model. As shown in Figure 5.1, InA is similar to LoRA as it optimizes rank decomposition matrices while keeping the pre-trained weights frozen. InA gates the passing information from the "internal" aspect, namely, by setting one threshold to control the passing information. However, LoRA gates the passing information from the 'external' aspect, that is, compressing the information using a low rank mechanism. Go further, InA introduces an additional threshold that weakens one part of the adaptation vector ($W_{inhibition}$). In the case of pre-trained language models, the inhibition vector is then used to reweigh irrelevant knowledge while retaining useful information through the non-inhibited part.

Figure 1.4 illustrates a practical example demonstrating the effectiveness of the proposed Inhibition Adaption in eliminating answer-irrelevant parts of the intrinsic rank, such as 'I' and 'My'. We hypothesize that the distribution of this intrinsic rank resembles a Gaussian-like distribution with a concentrated center and two sparse tails. In order to reduce the influence of task-irrelevant features during fine-tuning of pre-trained LMs, InA removes one tail by subtracting a proper threshold. The contributions of InA are as follows:

(a) InA effectively inhibits irrelevant information during fine-tuning on downstream tasks like GLUE and SQuAD, enabling the model to focus more on task-related information and eliminating the impact of irrelevant knowledge.

(b) InA benefits from proper activation functions with relatively flat negative tails. GeLU or LeakyReLU, which have small negative tails, outperform other activation functions like ReLU. SELU and ELU, with long and upturned tails, do not perform as well with InA.

(c) InA shares the same trainable parameter with LoRA, enabling it to inherit the knowledge compression ability from LoRA. Additionally, InA gains the capability to sup-

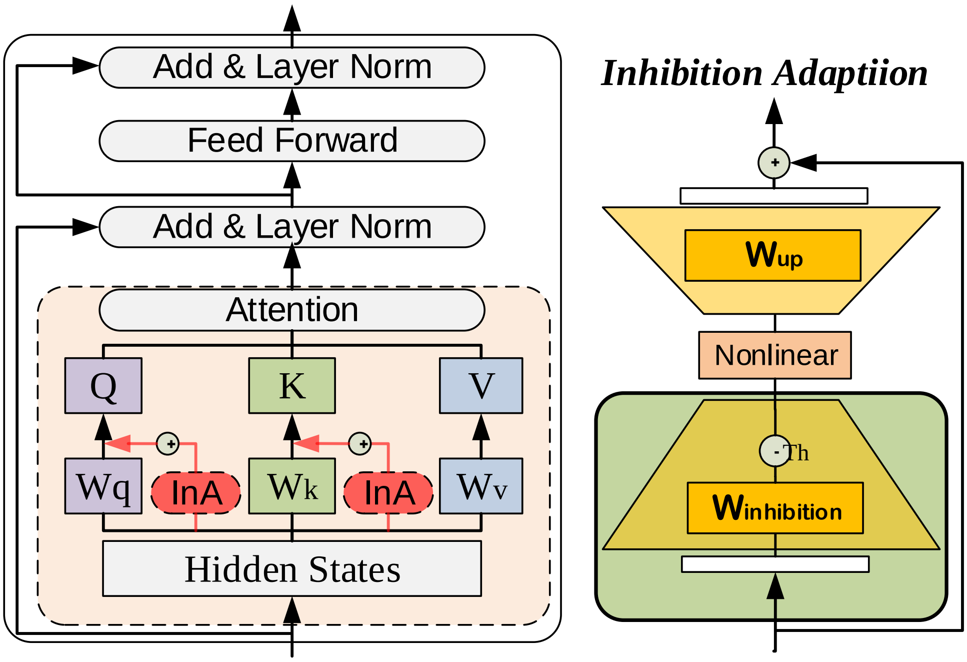


Figure 5.1: Illustration of the transformer architecture and our proposed parameter-efficient tuning method: inhibition adaption.

press task-irrelevant knowledge by subtracting a threshold.

5.2 Problem Statement

In the prior work of LoRA [52], authors only used the similarity matrix to compare the difference between LoRA fine-tuning and fully fine-tuning methods. There is no straight forward visualization result that can show us which part has been tuned by such methods. In addition, when using LoRA fine-tuning method on LMs, we found that although the low rank "bottleneck" can compress information and reweight the pre-trained parameters, such compressed information always contains noise and task-irrelevant knowledge. As shown in Figure 1.4, we present an example: input = ['I put my red bag in the black bag. What is the colour of my bag ?'], target = ['red']. When the threshold is 0, InA will become to LoRA, as InA also uses low rank to compress the passing information. The target-irrelevant knowledge in this case includes pronouns (e.g., I, my, and what), nouns (e.g., bag), verbs (e.g., put), definite articles (e.g., the), and adjectives (e.g., black and colour). Both full FT and adaption FT methods still retain this target-irrelevant information, which can distract the model from focusing on the actual target knowledge. When the target is specified as ['red colour'], the relevant knowledge should be the adjective "colour." Figure 1.4 is a cross attention map, and it presents the "word connection" between the column and the row word lists. The "word connection" between “I” and “red” is reasonable, but the most important “word” should be “red”. To make attention layers pay more attention to most important “words”, that means making attention layers more concentrated, the noise words, such as “I” should be inhibited. Therefore, it is essential to eliminate such target-irrelevant information to ensure the model’s output is more concentrated on the desired target. On the right-hand side of Figure 1.4, InA is introduced as a method to reduce the influence of the target-irrelevant knowledge, such as the pronoun ”I.”

Figure 1.4 shows a practical example using InA in the $BERT_{large}$ model, which has been fine-tuned under question-answering datasets. Left panel explains the potential risk of LoRA, and right panel presents the visualization of the attention score on last attention layer based on prior work [202]. The text is ’**I put my red bag in the black bag.**’, and the question is ’**What is the colour of my bag?**’, Therefore, the answer should be ’**red**’. There are two colours: red and black. Classical fine-tuning and adaption fine-tuning methods, such as LoRA, on downstream NLU tasks tend to choose the proper features from the entire ’redundant’ feature pool. This cannot essentially eliminate the influence of task-irrelevant words, for example, ’**I**’ and ’**My**’. After five epochs of InA fine-tuning, our inhibition vector can learn an incomplete intrinsic rank whose sole tail was eliminated by InA. Finally, activated by GeLU, which has a small negative tail, this incomplete intrinsic rank can provide the pre-trained weights with a small negative vector. Thus, these answer-irrelevant parts—’**I**’ and ’**My**’—in the intrinsic rank will be weakened or eliminated (see red stars in the right panel). We finally conclude that after InA fine-tuning, attention layers will pay less attention to such task-irrelevant information.

5.3 Explanation of Shunting Inhibition

5.3.1 Shunting Inhibition (Gate with Inhibition)

The design of a gated structure with inhibition draws inspiration from the shunting inhibition mechanism ([1], [203], [204]). The left panel in Figure 5.2 illustrates how shunting inhibition works, with its on (the red box) and off (the green box) states. When the gate of shunting inhibition is off, the signal transmission occurs across the joint, which can be influenced by shunting synapses. These shunting synapses play a crucial role in regulating neuronal function, and their activation can affect signal reception and transmission. In the context of ANNs, shunting can be described and interpreted as a gating mechanism in most articles, but researchers have often overlooked the inhibitory mechanism in the past. Shunting inhibition employs the shunting mechanism to select active neuron units, with its primary function being the selection, weakening, or strengthening of quantized features in ANNs.

In contrast to excitatory synapses, certain neurotransmitter-gated ion channels can direct the postsynaptic potential towards the resting potential or inhibit the effects of excitatory synapses ([203]). Such synapses are collectively referred to as 'inhibitory'. An example of inhibitory synapses involves the neurotransmitter $GABA$, which has both a fast receptor known as $GABA_A$ and a slower receptor called $GABA_B$. Additionally, the neurotransmitter dopamine has several receptor types, some of which are excitatory and some inhibitory. Inhibition can be subtractive, as it reduces the membrane potential, or divisive, as it modulates the effect of excitation. For instance, $GABA_A$ receptors have no effect on the membrane potential when it is at rest, so they do not further reduce the potential. Inhibitory synapses located close to the cell body can have modulatory (multiplicative) effects on the summed Excitatory Postsynaptic Potentials (EPSPs).

5.3.2 Membrane Potentials and Threshold

In Figure 5.2, the right panel illustrates the rationale behind setting the threshold between 10% and 30%. The red line represents the threshold for inhibition, and the membrane potentials typically range from $-70mV$ to $+30mV$. Considering the inactivated range of membrane potentials, we choose a threshold of approximately 15% (within the range of 10% to 30%). Not all neurons can act in the same way, and some may have a lower threshold of 1–5%. When the voltage exceeds the threshold, depolarization occurs following the activation. We assume that the distribution of activated features in artificial neural networks follows a Gaussian-like pattern. Commonly used activation functions like Softmax (Softmax), Tanh (Tanh), Rectified Linear Unit (ReLU) ([205]), Parametrised ReLU ([206]), Exponential Linear Unit (ELU) ([207]), Self-Gated Activation Function (Switsh) ([208]), Gaussian Error Linear Unit (GeLU) ([209]), and Scaled Exponential Linear Unit (SELU) ([210]) directly activate all features. However, to avoid the influence of unimportant features, those whose activated values fall below the threshold should be inhibited. These features have little significance for specific tasks, as the pre-trained model already provides highly quantized features for downstream fine-tuning tasks.

5.4 Related Work

5.4.1 Transformer-based language models

Heavily relying on the self-attention mechanism, Transformer ([211]), a sequence-to-sequence architecture, has dominated NLP and become SOTA for many tasks. Exploring the mechanism of scaling Transformer (by scaling model size, dataset size, model

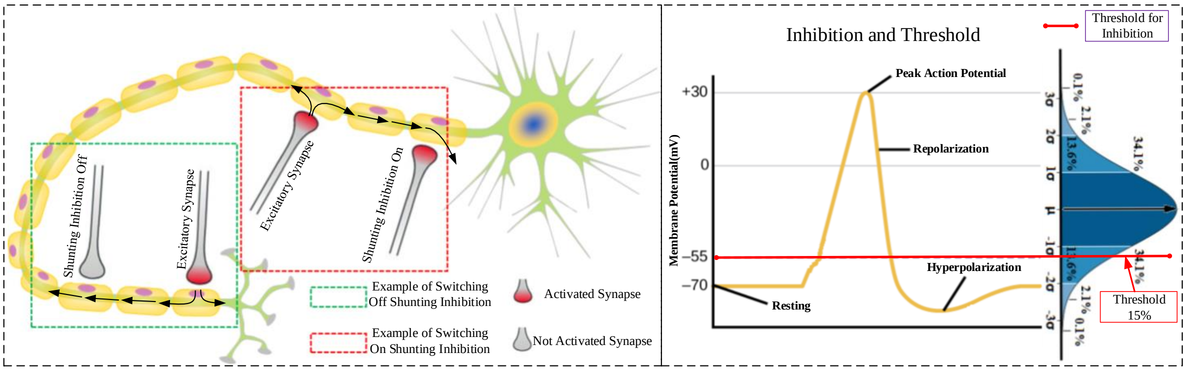


Figure 5.2: Inspiration from Neuroscience: Gate With Inhibition.

shape, context length and batch size), which was encouraged by the scaling law ([212]), has accelerated the capacity of various language models, such as BERT ([213]–[216]), RoBERTa ([217]), A Lite BERT (ALBERT) ([218]), DeBERTa ([219], [220]), sparse Switch-Transformer-1.6T ([221]) and Swin-Transformer ([222], [223]). Over the years, the capacity of language models has seen a dramatic thousandfold improvement. Taking single-head attention as an example, the self-attention operation with bias can be formulated as ([211], [216], [219], [222]):

$$Q = HW_q + b_q, K = HW_k + b_k, V = HW_v + b_v \tag{5.1}$$

$$A = \frac{QK^T}{\sqrt{D}}, \tag{5.2}$$

$$H_o = softmax(A + b_a)V, \tag{5.3}$$

where $H \in R^{M\times d}$ represents the input hidden vectors; $H_o \in R^{M\times d}$ is the output of the self-attention; $Q, K, V \in R^{M\times d}$ are the $Query, Key$ and $Value$ matrices; $W_q, W_k, W_v \in R^{d\times d}$ are the projection matrices; $A \in R^{M\times d}$ is the attention matrix; M is the length of the input sequence; $b_a, b_q, b_k, b_v \in R^{M\times M}$ are the relative position bias terms for each head and D is the dimension of hidden states.

5.4.2 Fine-tuning on NLP downstream tasks

Current SOTA systems for NLP tasks are based on the fine-tuning of pre-trained LMs. Classical fine-tuning methods should retrain the pre-trained model, which has been trained on general domains to fit a specific task ([213]). To maximise the downstream performance, variants of the vanilla Transformer (e.g., merely learning a subset of the parameters) compel practitioners to retrain all LM weights. For other conditional NLP tasks, such as question answering and dialogue generation, fine-tuning is also the prevalent paradigm ([219], [224]). In this paper, we focus on text classification, question answering and text adversarial generation tasks, as well as the three most frequently used pre-trained LMs: BERT, RoBERTa and DeBERTa. However, due to the large checkpoint and the high hardware barrier to entry, the enormity of these pre-trained LMs makes it challenging to perform fine-tuning in the usual way.

5.4.3 Parameter-Efficient Fine-Tuning

Adapters Tuning. The adapter tuning mechanism inserts several vectors (adapters) between transformer layers ([56]). The adapter module uses two projections, $W_{down} \in R^{d\times k}$ and $W_{up} \in R^{k\times d}$, first, to project H_o to a lower-dimensional space specified by the bottleneck dimension k, which is followed by a nonlinear activation function $f(\cdot)$, second, to project the computed product back with up-projection W_{up}. The final output of H_o after using adapters is:

$$H_o \leftarrow H_o + f(HW_{down})W_{up}. \tag{5.4}$$

One more efficient adapter variant ([225]) has been proposed, and it is inserted a Forward Neural Network (FNN) only after the 'add and layer norm' sub-layer.

Prefix and Infix Tuning. Prefix tuning prepends l tunable prefix vectors to the keys and values of the multi-head attention on every layer ([53]). By respectively concatenating or inserting two prefix vectors, $P_k \in R^{M\times p}$ and $P_v \in R^{M\times p}$ (p is the length of the inserted vector), to the head or middle of the original projection matrices K and V, new prefixed or infixed $Keys$ and $Values$ in the multi-head attention can be formed as:

$$W_k^{(i)} : prefix = concat(P_k^{(i)}, CW_k^{(i)}), \tag{5.5}$$
$$W_v^{(i)} : prefix = concat(P_v^{(i)}, CW_v^{(i)}), \tag{5.6}$$
$$W_k^{(i)} : infix = insert(CW_k^{(i)}, I_k^{(i)}), and \tag{5.7}$$
$$W_v^{(i)} : infix = insert(CW_v^{(i)}, I_v^{(i)}). \tag{5.8}$$

Given a sequence of m vectors, $C \in R^{M\times d}$, over which we would like to perform attention, multi-head attention performs the attention function in parallel on N_h heads. P_k, P_v, I_k and I_v are respectively split into N_h head vectors. $P_k^{(i)}$, $P_v^{(i)}$, $I_k^{(i)}$ and $I_v^{(i)} \in R^{M\times p}$ denote the i-th head vector. $W_k^{(i)}$ and $W_v^{(i)} \in R^{M\times(p+d)}$ denote the i-th prefix (or infix) head vector.

LoRA Tuning. LoRA injects trainable low-rank matrices into transformer layers to approximate the weight updates ([52]). By using a low-rank decomposition $W_0 + \Delta = W_0 + BA$, where B and A is respectively $W_{down} \in R^{d\times r}$ and $W_{up} \in R^{r\times k}$, LoRA updates the $query$ and $value$ projection matrices (W_q, W_v) in the multi-head attention sub-layer.

For the specific hidden input H, LoRA modifies the projection output H_o as:

$$H_o \leftarrow H_o + s \cdot f(HW_{down})W_{up}, \tag{5.9}$$

where $s \geqslant 1$ is a tunable scalar hyperparameter.

Others. Other parameter-efficient tuning methods include BitFit ([226]) which only fine-tunes bias vectors in the pre-trained model, diff-pruning ([227]) which learns a sparse parameter update vector, GLoRA ([228]) which generalizes the LoRA and QLoRA ([229]) which quantizes the LoRA with 4 or 8 bits.

5.4.4 Threshold and Inhibition

The threshold mechanism has been mostly used in deep SNNs ([201], [230]). A higher threshold will prevent the neuron from firing ('dead-neuron' problem), and a lower threshold will cause excessive firing. Both affect the ability of the neuron to differentiate between these two input patterns ([231]). The firing thresholds are also fixed ([232]) or selected based on some heuristics ([230], [233]). The threshold was selected as the maximum preactivation of each layer in [230]. [233] selected a certain percentile of the preactivation distribution as the threshold. Some recent works employ leak/threshold optimisation, but their application is limited to simple datasets ([234]). Most of these articles applied a threshold to SNNs, but they are facing the challenge of proposing improper methods of selecting the membrane leak and the threshold. To our best knowledge, there is no example of applying inhibition to a Transformer architecture.

5.5 Inhibition Adaption

InA, consists of a stack of gate blocks with an additional inhibition. The gate block can determine which features should be focused on, and inhibition can control the opening level of the gate. In this article, we use different inhibition-level percentiles Inh_p on different downstream tasks. Specifically, we set $Inh_p = 0.3$ when fine-tuning text classification and $Inh_p = 0.9$ when fine-tuning question answering and text adversarial generation tasks. In Figure 1.4, there is an example which illustrates how InA works and how it can inhibit the attention score when fine-tuning the question-answering task.

5.5.1 Inhibited Adaption

InA also inserts trainable inhibition matrices into transformer layers to approximate the weight updates. By using a low-rank decomposition $W_0 + \Delta = W_0 + W_{down}$, where

$W_{down} \in R^{d\times r}$, $W_{up} \in R^{r\times k}$, $Th \in R^{1\times r}$, InA updates the $Query$ and Key projection matrices (W_q, W_k) in the multi-head attention sub-layer. For the specific input H, InA modifies the projection output H_o as:

$$H_o \leftarrow H_o + s \cdot f(HW_{down} - Th)W_{up}, \tag{5.10}$$

where $s \in \{0, 1\}$ is a tunable scalar hyperparameter, and Th is the threshold.

Notation. We denote input hidden vectors as $H \in R^{M\times d}$ and the output of self-attention as $\bar{H}_o \in R^{M\times d}$. $W_k, W_q, W_v \in R^{d\times d}$ are the projection matrices.

Motivation. The motivation of InA on Transformer is to assemble a flexible gate with an adjustable inhibition vector to fine-tune downstream tasks. In addition, it should be able to automatically learn to rarefy tense features without sparsity settings. Under transfer learning, pre-trained language models can provide features for downstream tasks. The inhibition vector with a gate mechanism can learn to adjust and inhibit the provided features, and it finally makes tunable weights fit into a specific downstream task by fine-tuning. We formulate the linear InA layer as:

$$I_k = f(HW_{k\_down} - Th_k)W_{k\_up}, \tag{5.11}$$

$$I_q = f(HW_{q\_down} - Th_q)W_{q\_up}, \tag{5.12}$$

where $I_k \in R^{M\times d}$ and $I_q \in R^{M\times d}$, respectively, is the $Inhibition$ matrix in Key side and $Query$ side; f is the activation function; Th_k is the product of $max(HW_{k\_down}) \times Inh_p$ and Th_q is the product of $max(HW_{q\_down}) \times Inh_p$.

5.5.2 Inserting InA into Transformer

How shall we further adjust the adaptivity of LMs? And how do we select befitting features in such a huge feature pool after the pre-training? By using subtraction (- threshold Th_q), we propose Equation 5.11 and Equation 5.12, which have prejudice towards processing the features selection and can abandon features whose activated values are negative. With the use of inhibition, as shown in the right panel of Figure 1.4, the extra knowledge about '*I*' and '*my*' in the red box has been inhibited or removed. Under the application of the GeLU activation function, I_k and I_q will cut off the long negative tail to keep the concentrated features. This prejudice towards abandoning useless and counterproductive features will provide attention blocks with the ability to process dense features during fine-tuning.

The next step is to insert InA into Transformer attention blocks. Following the above

Table 5.1: Hyper-parameters for fine-tuning BERT, RoBERTa and DeBERTa with inhibited gate MLPs mechanism on down-streaming tasks.

Hyper-parameter	**BERT**(large)	**RoBERTa**(large)	**DeBERTa**(large)
Dropout of task layer	0.15	0.15	0.15
Warmup Steps	100	100	100
Learning Rates	5e-6	5e-6	5e-6
Batch Size	{16,32,64}	{16,32,64}	{16,32,64}
Weight Decay	0.01	0.01	0.01
Epochs	5	10	10
Learning Rate Decay	Linear	Linear	Linear
Optimizer	AdamW	AdamW	AdamW
Adam ϵ	1e-6	1e-6	1e-6
Adam (β_1, β_2)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Gradient Clipping	1.0	1.0	1.0
Inhibition Percentile	(0.0, 0.1, 0.3, 0.9)	(0.0, 0.1, 0.3, 0.9)	(0.0, 0.1, 0.3, 0.9)

elaboration, we formulate the linear InA on Transformer as:

$$V = HW_v + b_v, K = HW_k + b_k, Q = HW_q + b_q, \tag{5.13}$$

$$B_k = K + I_k, B_q = Q + I_q, \bar{A}_{kq} = \frac{B_q B_k^T}{\sqrt{D}}, \tag{5.14}$$

$$\bar{H}_o = softmax(\bar{A}_{kq} + b_{\bar{a}})V, \tag{5.15}$$

where $V \in R^{M \times d}$ is the $Value$ matrix; $B_k, B_q \in R^{M \times d}$ are respectively Key and $Query$ matrices with InA MLPs; $\bar{A}_{kq} \in R^{M \times d}$ is the attention matrix with InA MLPs and $b_{\bar{a}} \in R^{M \times M}$ is the relative position bias term for each head with InA MLPs.

Equations 5.13 and Equations 5.15 have the same form as the vanilla Transformer attention. They produce the Key, $Query$ and $Value$ projection matrices to represent the attribute of contexts. To select proper features that should be used to fit downstream tasks during fine-tuning, and to again modify the attribute of input contexts whose distribution should tend to fit target tasks, Equations 5.11 and Equations 5.12, on the one hand, utilise I_k and I_q to adjust projection matrices K and Q on a small-scale and, on the other hand, keep or enhance the important attributes of contexts relying on the addition between K and I_k (Q and I_q).

5.6 Experiments

5.6.1 Experiment Settings

Our experiments only depend on single-task fine-tuning. Our code is implemented based on the Huggingface Transformer ([235]). Following prior studies of language models ([52], [227]), we report results using large models. We use 8 × NVIDIA Tesla A100 with 40GB graphic memory cards to fine-tune the pre-trained models. Code and models are available at: `https://github.com/ChengKang520/gate-with-inhibition`.

5.6.2 Evaluation Datasets

This section evaluates the performance of InA in terms of downstream tasks on $BERT-large$ ([218]), $RoBERTa-large$ ([217]) and $DeBERTa-large$ ([219], [220]). Whether natural language understanding, question answering or generation, specifically, the benchmark GLUE ([236]), SQuAD v1.1 ([237]), SQuAD v2.0 ([237]) and SWAG ([238]), we followed the adapter fine-tuning setup ([225]) on $RoBERTa-large$ for a direct and fair comparison. Refer to Table 5.1 for detailed hyperparameters.

5.6.3 Fine-Tuning Implementation Details

Settings. Following BERT ([213]), RoBERTa ([217]) and DeBERTa ([219]), we adopt dynamic data batching. We also include span masking ([239]) as an additional masking strategy with a span size of up to three. For fine-tuning, we use Adam ([240]) as the optimiser for a fair comparison, and we train each task with a hyperparameter search procedure—each run takes about 1–2 hours on a DGX-2 node. All the hyperparameters are presented in Table 5.1. The model selection is based on the performance of the task-specific sets.

Our experiments are under fine-tuning on downstream tasks. Firstly, we set the inhibition percentile as 0% to test whether the result is similar to the settings without inhibited gate MLPs. Secondly, we set the inhibition percentile as 10% or 90% according to the performance of the first step. Finally, if the result, when the inhibition percentile is 10%, becomes better, we will set the inhibition percentile as 30%. If not, we will set the inhibition percentile as 90%.

5.6.4 Results

We summarise the efficiency performance of adaption FT methods and InA in Table 5.2. In addition to comparing with different adaption methods, by inserting InA into $BERT-large$, $RoBERTa-large$ and $DeBERTa-large$, we also summarise the results on eight NLU tasks of GLUE ([236]) in Table 5.3, as well as question answering – SQuAD v1.1 ([237]), SQuAD v2.0 ([241]) and Text Adversarial Generation: SWAG ([238]) in Table 5.4. In Table 5.5, we compare the performance of InA on the GLUE development set when fine-tuning $BERT-large$ with five epochs over five different activation functions. We also summarise the performance of different inhibition levels on these three large language models in Table 5.6.

Efficiency: Trainable Parameters and Speed

Additionally, we would like to discuss the efficiency gains of InA, such as the reduction in trainable parameters, and back-propagation speed and inference (complexity). We treat W_q (or W_k, W_v) as a single matrix of dimension $d \times d$. We denote the number of the prefix (resp. infix) tokens as l_p (resp. l_i). r is the low-rank mechanism that controls the bottleneck. In Table 5.2, the activation function of adapters and LoRA is ReLU; Prefix uses Softmax; and InA uses Leaky Rectified Linear Unit (LeakyReLU). Eventually, InA shows the fewest tunable parameters but the same inference complexity when using LeakyReLU. In Table 5.1, LeakyReLU has no obvious average gap with GeLU, because they almost have the same function and derivative waveform.

Table 5.2: The efficiency of InA and other adaptation FT methods in terms of trainable parameters, update speed (back-propagation) and inference (complexity).

Methods	Tunable Params	Inference	Update Speed
Fully FT	$T1 = 3 \times L \times d^2$	$T1$	$\mathcal{O}(2^n)$, GeLU
Adap FT	$T2 = 2 \times d \times r + r + d$	$T1 + T2$	$\mathcal{O}(n^2)$, ReLU
Prefix FT	$T3 = d \times (l_p + l_i)$	$T1 + T3$	$\mathcal{O}(2^n)$, Softmax
LoRA FT	$T4 = 2 \times d \times r$	$T1 + T4$	$\mathcal{O}(n^2)$, ReLU
InA FT	$T5 = 2 \times d \times r$	$T1 + T5$	$\mathcal{O}(n^2)$, LeakyReLU $\mathcal{O}(2^n)$, GeLU

5.6.5 Effectiveness: InA on Fine-tuning

Table 5.3: Comparison results of fine-tuning the GLUE development set on $BERT{-}large$, $RoBERTa{-}large$, $DeBERTaV2{-}large$ and $DeBERTaV3{-}large$ with InA (inhibition level percentile is 0.3). † indicates runs configured in a setup similar to [56] for a fair comparison.

Model-large & Method #Train	#Trainable Parameters	CoLA Mcc 8.5k	QQP Acc 364k	MNLI Acc 393k	SST2 Acc 67k	STS-B Corr 7k	QNLI Acc 108k	RTE Acc 2.5k	MRPC Acc 3.7k	Avg.
BERT [213]	336.0M	60.6	91.3	86.6	93.2	**90.0**	92.3	70.4	**88.0**	84.5
BERT [FT] †	336.0M	64.0	91.3	86.2	93.8	88.9	92.6	**71.4**	86.6	**84.35**
BERT [LoRA] †	0.8M	64.2±0.7	91.4±0.2	86.2±0.2	94.2±0.2	89.2±0.2	**92.7±0.1**	69.2±1.4	84.9±1.3	84.01
BERT [InA] †	0.4M	**65.9±0.6**	**91.5±0.1**	**86.3±0.2**	**94.4±0.2**	89.0±0.2	**92.7±0.1**	69.0±1.6	84.8±1.1	84.19
RoBERTa [217]	355.0M	68.0	**92.2**	90.2	96.4	92.4	93.9	**86.6**	**90.9**	88.82
RoBERTa [FT] †	355.0M	68.1	**92.2**	90.2	96.3	92.3	93.9	**86.6**	**90.9**	88.56
RoBERTa [Adpt]†[225]	0.8M	67.8±2.5	91.7±0.2	90.5±0.3	**96.6±0.2**	91.9±0.4	**94.8±0.3**	80.1±2.9	89.7±1.2	87.9
RoBERTa [Adpt]†[56]	0.8M	66.3±2.0	91.5±0.1	90.3±0.3	96.3±0.5	91.5±0.5	94.7±0.2	72.9±2.9	87.7±1.7	86.4
RoBERTa [LoRA]†[52]	0.8M	68.2±1.9	91.6±0.2	**90.6±0.2**	96.2±0.5	**92.3±0.5**	**94.8±0.3**	85.2±1.1	90.2±1.0	88.6
RoBERTa[InA] †	0.4M	**68.5±1.2**	**92.2±0.1**	90.2±0.4	96.4±0.3	92.0±0.3	94.4±0.4	85.2±0.7	90.8±0.5	**88.7**
DeBERTaV2 [219]	304.0M	70.5	92.3	91.1	96.8	92.8	95.2	88.3	91.9	90.00
DeBERTaV3 [220]	304.0M	75.3	93.0	**91.8**	**96.9**	93.0	96.0	**92.7**	92.2	**91.37**
DeBERTaV3 [FT] †	304.0M	74.3	93.0	91.0	96.2	92.6	95.4	90.3	90.7	90.44
DeBERTaV3 [LoRA] †	0.8M	75.6±1.2	93.1±0.1	91.0±0.2	96.6±0.3	92.8±0.2	96.0±0.1	91.2±0.7	92.9±0.2	91.15
DeBERTaV3 [InA] †	0.4M	**76.4±1.0**	**93.2±0.1**	90.9±0.3	96.6±0.4	**93.2±0.2**	**96.1±0.1**	90.7±0.8	**93.1±0.2**	91.28

Our settings for $BERT - large$ and $DeBERTa - large$ on InA are, respectively, similar to the input/output protocol for BERT ([213]) and DeBERTa ([220]) fine-tuning. Our settings for InA fine-tuning on $RoBERTa - large$ are, respectively, similar to the adaption fine-tuning method ([52], [225]).

Table 5.4: Comparison results of fine-tuning SQuAD v1.1, SQuAD v2.0 and SWAG on $BERT-large$, $RoBERTa-large$, $DeBERTaV2-large$ and $DeBERTaV3-large$ with InA (inhibition level percentile is 0.9). ⋆ indicates being run under the original configuration for a fair comparison. (Note that missing results in the literature are signified by '-').

Model-large & Method #Train	# Trainiable Parameters	SQuAD v1.1 F1/EM	SQuAD v2.0 F1/EM	SWAG Acc
BERT [213]	336.0M	90.9/84.5	**81.8/79.0**	**88.6**
BERT [FT] ⋆	336.0M	91.3/84.5	81.7/78.4	86.5
BERT [LoRA] ⋆	0.8M	91.3/84.5	81.7/78.4	86.5
BERT [InA] ⋆	0.4M	**91.3/84.6**	81.5/78.1	86.7
RoBERTa [217]	355.0M	94.5/88.9	**89.4/86.5**	89.9
RoBERTa [FT] ⋆	355.0M	94.1/88.4	88.9/86.0	88.9
RoBERTa [LoRA] ⋆	0.8M	94.4/88.7	88.8/86.0	88.9
RoBERTa [InA] ⋆	0.4M	**94.7/89.2**	89.1/86.3	88.9
DeBERTaV2 [219]	304.0M	**95.5/90.1**	90.7/88.0	90.8
DeBERTaV3 [220]	304.0M	-	91.5/89.0	**93.4**
DeBERTaV3 [FT] ⋆	304.0M	95.4/89.8	91.5/89.0	93.3
DeBERTaV3 [LoRA] ⋆	0.8M	95.3/89.9	91.5/89.0	93.2
DeBERTaV3 [InA] ⋆	0.4M	95.4/90.0	**91.6/89.0**	93.3

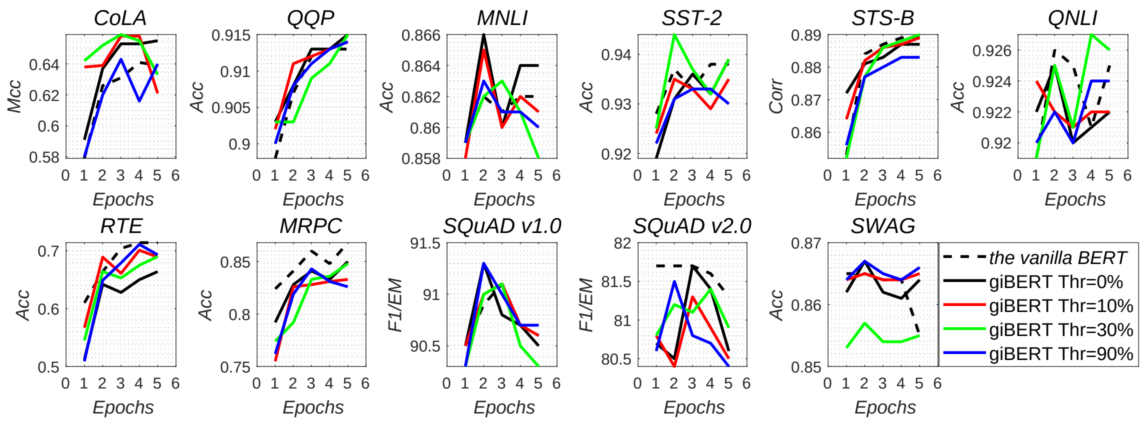


Figure 5.3: Plots of corresponding metrics according to the number of epochs on the validation split of GLUE, SQuAD v1.1, SQuAD v2.0 and SWAG. The giBERT means inserting InA (gate inhibition mechanism) into BERT.

5.6.6 InA on the Text Classification Task

We summarise the comparison results on these eight NLU tasks in Table 5.3 in terms of fine-tuning the architecture of inserting InA into the original $BERT-large$ $RoBERTa-large$ and $DeBERTa-large$. In Table 5.3, when using $BERT-large$ as the base, the average cannot catch up with the performance of using the classical FT method, but InA fine-tuning outperforms the classical FT method on six out of eight tasks. Although $RoBERTa-large$ with InA fine-tuning merely shows the fine-tuning advantage on Corpus of Linguistic Acceptability (CoLA), Quora Question Pairs (QQP) and Microsoft Research Paraphrase Corpus (MRPC) tasks, it can achieve the highest average result. Figure 5.5 shows the attention heatmap when using InA to fine-tune the GLUE tasks. Fine-tuning

$DeBERTaV3 - large$ with InA on GLUE can get five out of eight better results, even though it also cannot achieve a better average. From Table 5.3, we can find that when fine-tuning Recognizing Textual Entailment (RTE) and MRPC under InA, $BERT-large$ and $RoBERTa - large$ cannot always get a better performance than other FT methods. The inferred reason is that the extra tunable parameters cannot be efficiently fine-tuned with small data.

InA on the Question Answering Task

As we use three large language models as the baseline, $BERT-large$, $RoBERTa-large$ and $DeBERTa - large$, when fine-tuning with InA on SQuAD v1.1 and SQuAD v2.0 ([237]), we can find a weak improvement in Table 5.4. Moreover, the obviously dominant part is that InA inhibits the 'irrelevant knowledge' (e.g., 'I' and 'my') when $Inh_p = 0.9$ (See Figure 5.6). We infer that InA inhibits the information that has a relationship with the label (the label is 'red'), for example, the word 'my' in the phrase 'my red'. That is why InA can achieve relatively better results on the SQuAD task.

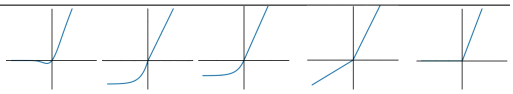
InA on the Multi-Choice Generation Task

In Table 5.4, for the SWAG text generation dataset ([238]), which introduces the task of grounded commonsense inference, unifying natural language inference and commonsense reasoning, we find there is no fine-tuning improvement. In Figure 5.10, the input is 'she opened the hood of the car'. Humans can reason about the situation and anticipate what might come next (the label is 'then, she examined the engine'). The inhibitor can reduce the influence of some information, but the reason why such 'unimportant knowledge' is required for the SWAG task is still not clear. We will perform more experiments to figure out the reason why InA cannot benefit SWAG in our future work.

Different Activation Functions on InA

We summarise the results of using different activation functions after setting the inhibition percentile at 30% in Table 5.5. When compared with other activation functions whose tails are zero or negative, the GeLU activation function, whose negative tails are short, achieves the best improvement of QQP, Stanford Sentiment Treebank (SST2), Stanford Question Answering Dataset (QNLI), MRPC and GLUE averages. Although LeakyReLU with a default slope gets outstanding performance on CoLA and RTE, the total effect on GLUE tasks is inferior to GeLU. LeakyReLU can provide more stable and smoother negative values, and this could be the reason why LeakyReLU can outperform GeLU on these two small downstream GLUE tasks. The negative value deriving from LeakyReLU

Table 5.5: When using different activation functions, we set the inhibition level percentile at 0.3 and present the comparison results on the GLUE development set within five epochs fine-tuning based on $BERT-large$.

Model-large #Train BERT(30%)	GeLU	SELU	ELU	LeakyReLU	ReLU
Functions					
CoLa (Mcc)	65.9	62.1	62.8	**66.6**	64.3
QQP (Acc)	**91.5**	63.2	63.2	91.4	91.4
MNLI (Acc)	86.3	35.4	35.5	86.3	86.3
SST2 (Acc)	**94.4**	50.9	92.9	93.6	93.1
STS-B (Corr)	89.0	32.0	77.0	88.9	**89.3**
QNLI (Acc)	**92.7**	50.5	92.0	92.3	92.3
RTE (Acc)	69.0	54.9	52.7	**70.0**	68.6
MRPC (Acc)	**84.8**	68.4	77.2	84.3	83.8
Avg.	**84.20**	44.41	69.15	84.18	83.64

activation would provide a stronger inhibition for BERT or variants of BERT (RoBERTa, DeBERTaV2 and DeBERTaV3). GeLU has a short and tender negative tail, and we eventually select it as the default activation function.

In Table 5.5, every activation function has its negative tail, except ReLU. Because the inhibition vector has subtracted one inhibition variable through the GeLU and LeakyReLU activation functions, some variables become negative, and the output of the inhibition layer at the end has more negative variables if setting Inh_p higher. Thus, we can slightly 'reweight' the Q and K matrices with this inhibition vector. The worse performance of SELU can be a contrary example because it has an upturned tail which provides bigger negative outputs.

Inhibition Level in InA

Table 5.6: Comparison results on fine-tuning the GLUE development set, SQuAD v1.1, SQuAD v2.0, and SWAG—Inserting InA into $BERT-large$(1*), $RoBERTa-large$(2*) and $DeBERTa-large$(3*). The values after each model are inhibition levels.

Model #Train		**GLUE**									**SQuAD v1.1**	**SQuAD v2.0**	**SWAG**
(Large Model on InA)		CoLA Mcc 8.5k	QQP Acc 364k	MNLI Acc 393k	SST2 Acc 67k	STS-B Corr 7k	QNLI Acc 108k	RTE Acc 2.5k	MRPC Acc 3.7k	Avg.	F1/EM 87.6k	F1/EM 130.3k	Acc 73.5k
1*	BERT(0)	65.5	91.5	86.6	93.9	88.7	92.5	66.4	85.0	83.76	91.1/84.3	***81.6/78.9***	86.6
	BERT(0.1)	65.8	91.4	86.5	93.5	88.9	92.4	70.1	83.1	83.96	91.1/84.4	81.3/78.5	86.5
	BERT(0.3)	***65.9***	***91.5***	86.3	***94.4***	89.0	***92.7***	69.0	84.8	84.19	91.1/84.4	81.4/78.1	***86.7***
	BERT(0.9)	64.3	91.4	86.3	93.3	88.3	92.4	71.1	84.3	83.70	***91.3/84.6***	81.5/78.1	86.7
2*	RoBERTa(0)	64.1	92.2	90.2	95.8	92.0	94.1	85.2	89.0	87.81	93.9/88.4	88.3/84.7	88.3
	RoBERTa(0.1)	65.5	92.0	89.5	95.6	92.4	94.4	83.4	***91.7***	88.05	94.1/88.8	88.5/85.5	88.4
	RoBERTa(0.3)	***68.5***	92.2	90.2	96.4	92.0	***94.4***	85.2	90.8	88.69	94.2/88.8	88.7/85.3	89.6
	RoBERTa(0.9)	67.5	92.1	89.6	95.8	91.6	94.1	85.2	89.7	88.20	***94.7/89.2***	89.1/86.3	89.9
3*	DeBERTaV3(0)	73.2	93.1	90.9	96.6	93.2	95.5	90.3	91.4	90.65	95.2/89.7	90.8/88.5	91.9
	DeBERTaV3(0.1)	**76.5**	**93.2**	90.8	96.2	**93.2**	96.0	90.0	92.3	91.03	95.3/89.9	91.2/88.7	93.3
	DeBERTaV3(0.3)	76.4	**93.2**	**90.9**	**96.6**	**93.2**	**96.1**	**90.7**	**93.1**	**91.28**	95.4/89.9	91.1/88.4	**93.5**
	DeBERTaV3(0.9)	72.8	93.0	90.9	96.2	92.6	95.5	89.5	90.7	90.19	95.4/90.0	**91.6/89.0**	93.3

We also summarise the performance of using four different inhibition levels in Table 5.6. For text classification tasks, when the inhibition level percentile is 0.3, InA can achieve the dominant results. In Figure 5.3, the inhibition mechanism affects the fine-tuning performance, especially when the inhibition level is between 10% and 30%. But for the question-answering and adversarial text-generation tasks, when the inhibition level percentile is 0.9, there is a weak improvement.

Trainable Weights by Using s on InA

InA on Single Key or Query Side. For the single side conditions (either on the Key or on the $Query$) and based on $DeBERTaV3-large$, we summarise the results in Table 5.7. When the inhibition level Inh_p is 0.3, we get the best GLUE average using InA both on the Key and on the $Query$. There are two unexpected findings when inserting InA into the single attention side (Key or $Query$). The first is that when setting the inhibition level $Inh_p = 0.0$, we can achieve the best result at 92.1% in terms of fine-tuning the RTE task. The second is that when fine-tuning the downstream SQuAD v1.1 task with 0.3 and 0.1 inhibition levels, the Key side and the Query side respectively present the best result at 95.8%/89.3% and 95.8%/89.5%.

Table 5.7: Comparison results on fine-tuning the GLUE development set, SQuAD v1.1, SQuAD v2.0, SWAG and NER. (Note that **Key*** and **Query*** respectively mean inserting InA into Transformers' Key side and Query side).

	Model #Train	**GLUE**									**SQuAD v1.1**	**SQuAD v2.0**	**SWAG**
	(Large)	CoLA Mcc 8.5k	QQP Acc 364k	MNLI-m/mm Acc 393k	SST2 Acc 67k	STS-B Corr 7k	QNLI Acc 108k	RTE Acc 2.5k	MRPC Acc 3.7k	Avg.	F1/EM 87.6k	F1/EM 130.3k	Acc 73.5k
Key*	giDeBERTaV3(0)	72.6	93.0	90.9/90.9	96.3	92.8	95.4	88.8	92.2	90.25	94.8/89.2	89.9/86.5	92.2
	giDeBERTaV3(0.1)	74.0	93.0	91.2/91.0	96.2	92.9	95.4	89.5	91.9	90.51	94.8/89.3	89.7/86.9	91.6
	giDeBERTaV3(0.3)	75.0	93.1	91.0/90.9	96.2	92.8	95.3	91.7	91.7	90.85	**<u>95.8/89.3</u>**	89.9/86.4	92.2
	giDeBERTaV3(0.9)	72.0	93.1	91.0/91.0	96.3	92.8	95.4	91.3	91.4	90.41	94.8/89.3	90.3/86.9	92.0
Query*	giDeBERTaV3(0)	71.9	93.0	91.0/90.9	96.2	92.8	95.3	92.1	90.2	90.31	94.7/89.2	90.1/86.9	92.2
	giDeBERTaV3(0.1)	73.2	92.9	91.3/90.9	96.3	92.7	95.1	89.2	90.2	90.11	**<u>95.8/89.5</u>**	90.4/87.7	92.2
	giDeBERTaV3(0.3)	73.5	92.9	91.3/90.9	96.2	93.0	95.4	89.5	91.9	90.46	94.8/89.3	89.7/86.9	91.6
	giDeBERTaV3(0.9)	74.2	93.0	90.8/90.8	95.6	92.9	95.4	90.6	90.2	90.34	94.8/89.5	89.8/86.7	92.0

Inserting InA into Several Last Layers. To find the best inserting position, for example, which layer in BERT-like architectures needs inhibition, as well as ascertain how deep the inhibition should be set (for example, from the 16_{th} layer to the 24_{th} layer), we summarise the relevant results in Table 5.8 based on $DeBERTaV3-large$. We roughly disassemble the DeBERTa architecture in Figure 5.4 and, depending on this, we insert InA into several last layers (last 3, 6 and 12 layers).

5.7 Analysis and Discussion

We now proceed to empirically validate the effectiveness of InA. Based on experimental results of the benchmarks, we address and answer the following three questions: **Q1:** Should we really need inhibition during Adaptation fine-tuning? And how does the InA method work during fine-tuning? **Q2:** If we need it, how to choose the inhibition level

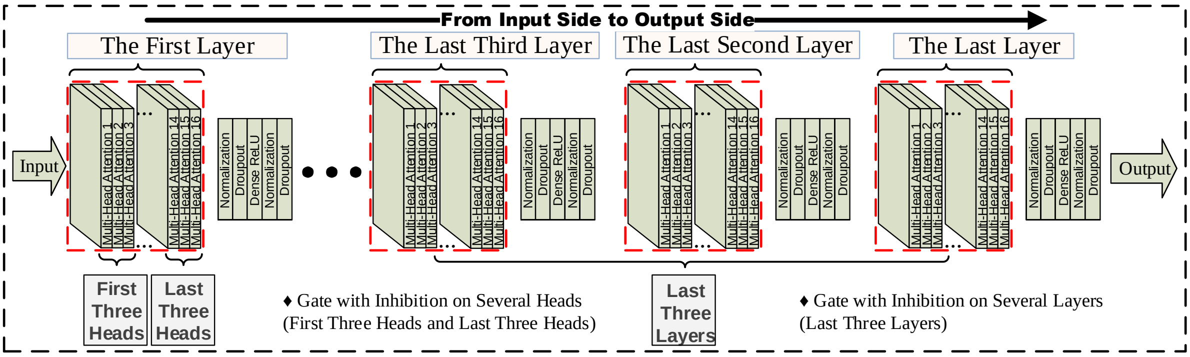


Figure 5.4: Roughly disassembled DeBERTaV3 architecture.

Table 5.8: Comparison results on fine-tuning the GLUE development set, SQuAD v1.1, SQuAD v2.0, SWAG and NER on language models' several last layers.

Model #Train		GLUE									SQuAD v1.1	SQuAD v2.0	SWAG
(Large Model on InA)		CoLA Mcc 8.5k	QQP Acc 364k	MNLI Acc 393k	SST2 Acc 67k	STS-B Corr 7k	QNLI Acc 108k	RTE Acc 2.5k	MRPC Acc 3.7k	Avg.	F1/EM 87.6k	F1/EM 130.3k	Acc 73.5k
Last 3	DeBERTaV3(0)	73.5	92.9	91.0	96.6	92.8	95.5	89.2	90.7	90.27	94.7/89.1	89.7/86.9	91.4
	DeBERTaV3(0.1)	73.2	93.0	90.9	96.5	92.9	95.8	90.6	91.1	90.50	94.3/88.6	89.5/86.1	91.0
	DeBERTaV3(0.3)	74.2	93.0	91.1	96.2	93.0	95.3	90.2	91.4	90.55	94.6/89.1	89.7/86.8	91.3
	DeBERTaV3(0.9)	74.4	93.0	90.9	96.0	93.0	95.3	89.5	91.7	90.48	94.2/88.5	89.9/86.9	91.2
Last 6	DeBERTaV3(0)	72.6	93.0	91.1	96.2	92.9	95.3	88.8	90.9	90.10	94.5/89.2	89.5/86.8	91.2
	DeBERTaV3(0.1)	72.9	93.0	91.1	96.2	92.9	95.3	88.8	90.9	90.14	94.5/88.9	89.5/86.7	91.3
	DeBERTaV3(0.3)	73.6	93.2	91.0	96.3	93.0	95.7	88.1	91.2	90.26	94.6/89.1	89.5/86.7	91.3
	DeBERTaV3(0.9)	74.2	93.1	90.9	96.0	93.0	95.4	88.5	90.9	90.25	94.7/89.0	89.5/86.8	91.2
Last 12	DeBERTaV3(0)	73.4	93.0	91.0	96.2	92.9	95.3	89.2	90.9	90.24	94.5/89.0	89.4/86.7	91.2
	DeBERTaV3(0.1)	73.9	93.0	91.0	96.2	92.9	95.5	89.9	91.1	90.44	94.4/88.9	89.5/86.9	91.2
	DeBERTaV3(0.3)	74.8	93.2	91.0	96.3	93.0	95.6	89.8	91.3	90.63	94.6/89.0	89.5/86.8	91.3
	DeBERTaV3(0.9)	74.2	93.1	90.9	96.0	93.0	95.3	89.3	90.9	90.34	94.7/89.0	89.4/86.7	91.2

Inh_p and select a good rank s in real cases? **Q3:** Dose the inhibition adaptation matrix $W_{inhibition}$ really inhibit irrelevant knowledge? If yes, which irrelevant knowledge will be inhibited in practice? We believe that our answers to **Q2** and **Q3** shed light on the fundamental principles of using pre-trained language models on downstream tasks.

5.7.1 Difference Between LoRA and InA

We conducted experiments to ensure a fair comparison with LoRA. From Figure 5.5a) to Figure 5.10a), when inhibition level is 0, namely, when InA is initialized as LoRA, InA can reweight the pre-trained parameters. However, if InA sets a higher inhibition lever, such as $Inh_p = 0.3$, as seen from Figure 5.5c) to Figure 5.10c), InA presents the ability to further adapt the activated features to weaken the influence of the irrelevant information. A lower Th has a weaker influence on the inhibition of passing information, but a higher one will inhibit most passing information. Although the performance between LoRA and InA is quite similar, InA still has the advantage to inhibit the passing information by using a proper Th. InA not only inherits the ability of LoRA to compress the passing

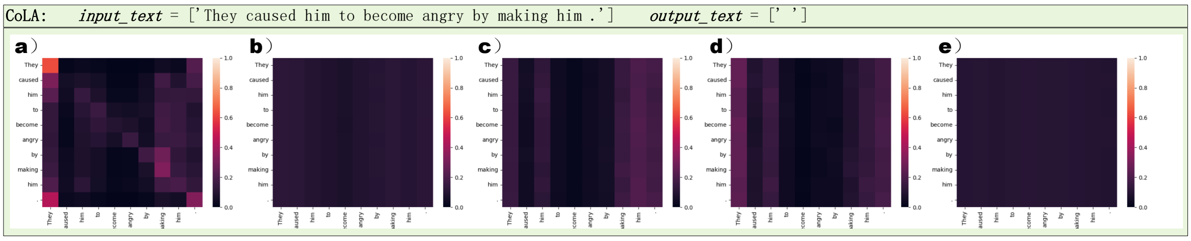


Figure 5.5: From left to right, fine-tuning $BERT-large$ on CoLA with **a)** no-InA, **b)** InA(0.0), **c)** InA(0.1), **d)** InA(0.3), **d)** InA(0.9).

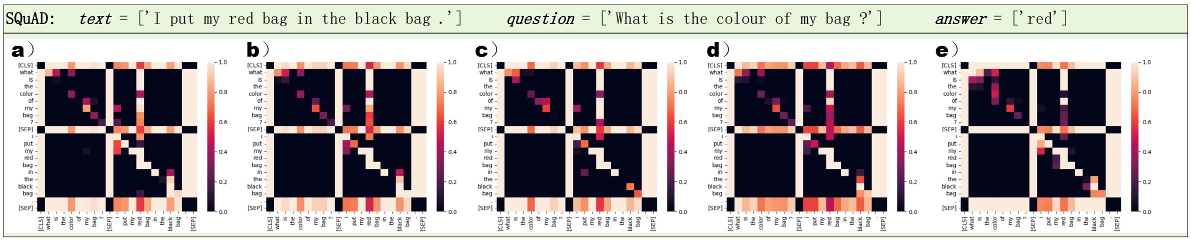


Figure 5.6: From left to right, fine-tuning $BERT-large$ on SQuAD with **a)** no-InA, **b)** InA(0.0), **c)** InA(0.1), **d)** InA(0.3), **e)** InA(0.9).

information but also inhibit the passing information by using a threshold. InA offers two key advantages over adapters like LoRA and Adapter: **(1)** InA incorporates the rank of the adapter r to control redundant information flow through the bottleneck. This allows the passing information to be treated as compressed compared to the original information in LoRA; **(2)** InA also utilizes a subtracted threshold to reduce the passing information, effectively controlling the Inhibition threshold. This achieves the same effect as adjusting only r. The passing information in InA can be considered as incomplete, as it discards task-irrelevant parts of the original information.

5.7.2 Should we need inhibition during fine-tuning? And how does it work?

Redundant features that we obtained from pre-trained language models can reduce the performance when using the full fine-tuning method, especially when fine-tuning on a small dataset. Therefore, we apply a similar MLP architecture (as the one used in gate multilayer perceptron (gMLP) [242]) with the proposed inhibition mechanism to address this challenge, and it eventually shows a positive effect on reducing the irrelevant knowledge. We need InA when fine-tuning pre-trained LMs on downstream NLU tasks.

Because RoBERTa has pre-trained over 160GB texts with a larger mini-batch and a larger byte-level of Byte-Pair Encoding [243]. This finally prompts RoBERTa to gain a robust speciality - the capacity to handle large and wide vocabularies [217]. InA on

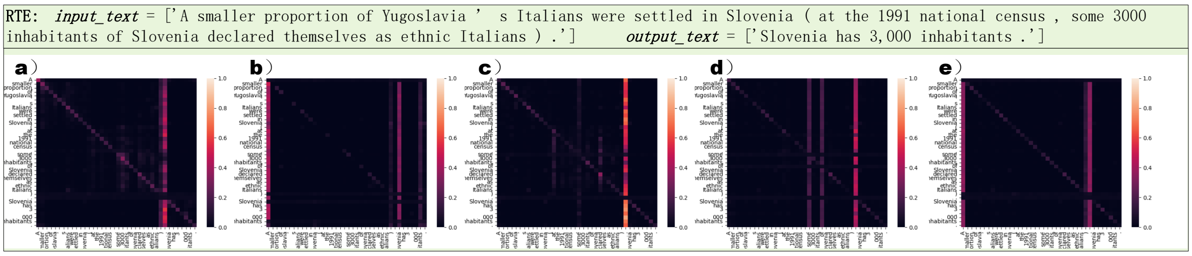


Figure 5.7: From left to right, fine-tuning $BERT-large$ on RTE with **a)** no-InA, **b)** InA(0.0), **c)** InA(0.1), **d)** InA(0.3), **d)** InA(0.9).

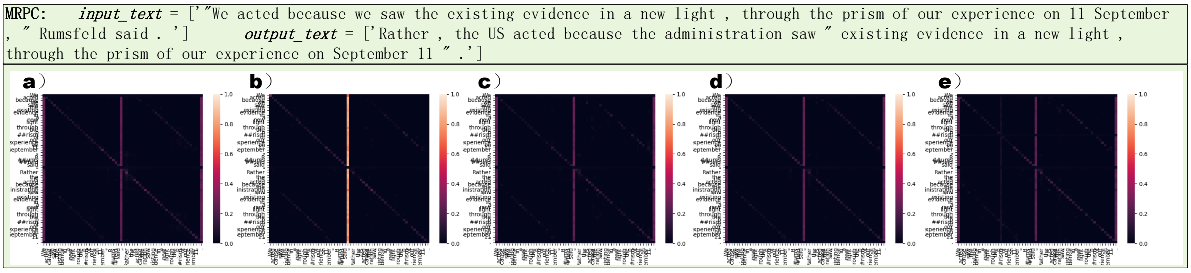


Figure 5.8: From left to right, fine-tuning $BERT-large$ on MRPC with **a)** no-InA, **b)** InA(0.0), **c)** InA(0.1), **d)** InA(0.3), **e)** InA(0.9).

RoBERTa cannot achieve better results on RTE, and we infer that InA also needs more fine-tuning steps to scale the robust large weights on small downstream tasks. DeBERTa has two special vectors inside: the content vector and the position vector. Its attention weights among words are computed by using disentangled matrices respectively based on their content and relative position vectors [220], and this makes a stronger contextual connection among input word vectors. InA on DeBERTa can inhibit the redundant contextual connection among input word vectors by scaling these disentangled matrices. In other words, these inhibited gate MLPs can also act as a sparse layer which provides DeBERTa with positive weights to concentrate more on significant connections.

5.7.3 How to choose the inhibition level Inh_p and select a good rank s in real cases?

We further investigate the influence of Inh_p on fine-tuning GLUE, SQuAD and SWAG tasks. From Table 5.6, in terms of the overall performance, we find that a proper inhibition level (e.g., $Inh_p = 0.3$) can make the text classification results better, and a strong inhibition (e.g., $Inh_p = 0.9$) can benefit the question-answering task. In practice, we find that if the size of the downstream dataset is small (e.g., RTE), it is better to insert InA into a $Query$ with 0% inhibition or insert InA into double sides ($Query$ and Key) with an inhibition level of 30%. To our best knowledge, we conclude the experience about how

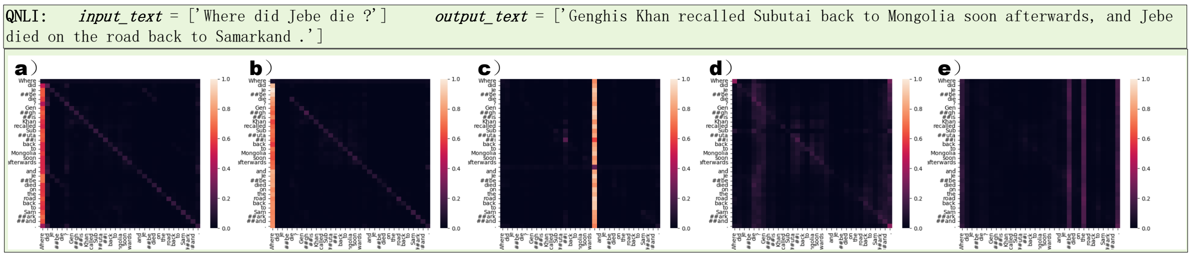


Figure 5.9: From left to right, fine-tuning $BERT-large$ on QNLI with **a)** no-InA, **b)** InA(0.0), **c)** InA(0.1), **d)** InA(0.3), **e)** InA(0.9).

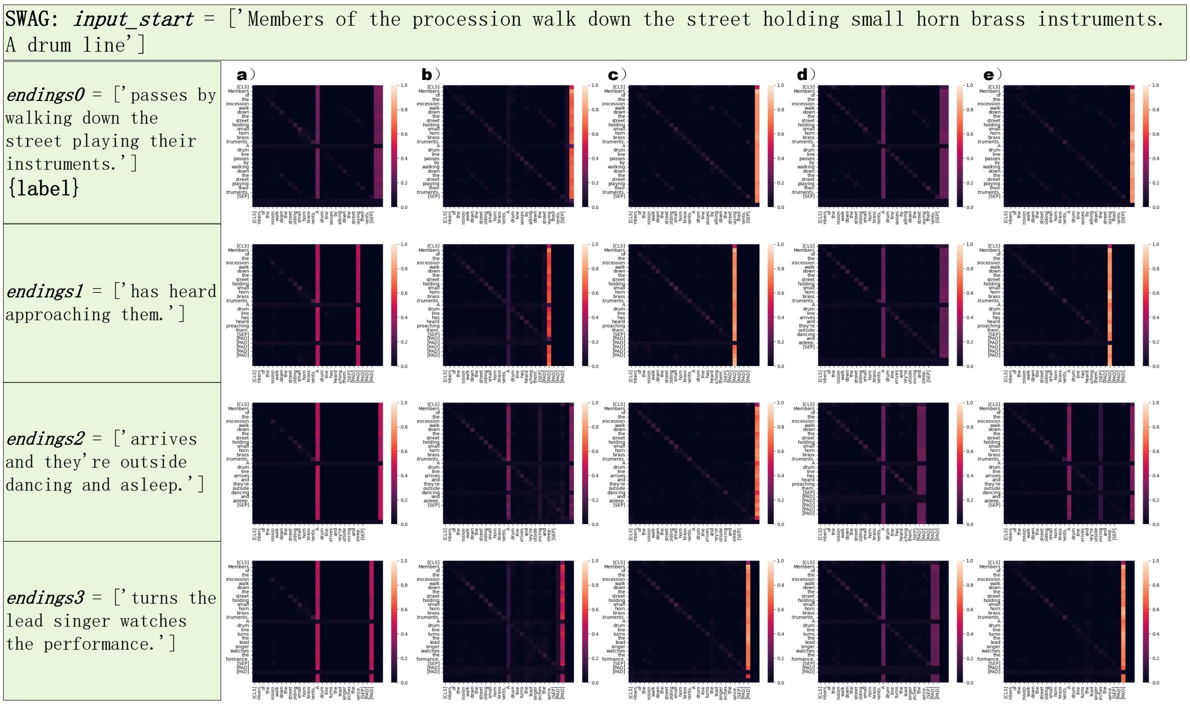


Figure 5.10: From left to right, fine-tuning $BERT-large$ on SWAG with **a)** no-InA, **b)** InA(0.0), **c)** InA(0.1), **d)** InA(0.3), **e)** InA(0.9).

to find a proper inhibition threshold during the InA fine-tuning as: **(1)** firstly use the 0% inhibition, **(2)** if the result is better than that without InA, choose the inhibition threshold between 10% and 30%, **(3)** and on the contrary, use a higher inhibition threshold (e.g., 90%). In order to find a good rank s in practical cases, we summarize the results of inserting InA into several last Transformer attention layers in Table 5.8. We find that inserting InA into several last layers cannot get any obvious improvements when fine-tuning pre-trained DeBERTaV3 on downstream tasks. However, the useful finding is when fine-tuning on downstream NLU tasks, we should insert InA into all layers or as many layers as when the memory is available.

5.7.4 Can InA really inhibit irrelevant knowledge? How can InA inhibit them?

To answer these two questions, we finally turn our attention to the performance of inhibition vector $W_{inhibition}$ and focus on its function to inhibit irrelevant knowledge.

For the first question, when fine-tuning on the SQuAD task under five conditions: without InA, with InA when $Inh_p = 0.0, 0.1, 0.3, 0.9$, we plot the averaged attention score heatmap of the last layer (we average all heads in the 24-th layer) in Figure 5.6 As the inhibition level is becoming stronger (from $Inh_p = 0.0$ to $Inh_p = 0.9$), the attention score of "I" and "my" is gradually reduced, which means the influence of such irrelevant knowledge is eventually eliminated, while the overall distribution trend of the attention score keeps the same. In Figure 5.6, the $text$ = ['I put my red bag in the black bag .'], and the $question$ = ['What is the colour of my bag ?'], then the $answer$ = ['red']. After the fine-tuning process, we expect that the concentrated words of the question vector should be "colour", and the concentrated words of the text vector should be "red" or "black". However, the classical FT method presents a residual problem with the "noise" knowledge. InA finally eliminates the "noise" knowledge as the Inh_p grows higher (e.g., "I" and "my" in Figure 5.6 **a)** and **e)**). This brings us the answer that InA can inhibit irrelevant knowledge during fine-tuning.

For the second question, according to how much it can tune the attention scores, we also plot averaged attention score heatmaps over five downstream tasks: CoLA (seen in Figure 5.5, RTE (seen in Figure 5.7), MRPC (seen in Figure 5.8), QNLI (seen in Figure 5.9) and SWAG (seen in Figure 5.10). From Table 5.3, when using InA, we can find an obvious improvement in CoLA compared with other results. The Corpus of Linguistic Acceptability of 'They caused him to become angry by making him .' is $False$. In Figure 5.5 **a)**, the attention block mostly concentrates on ['They', 'him', 'to', 'by', 'making', 'him' '.']. After the use of InA fine-tuning, the difference between Figure 5.5 **a)** and Figure 5.5 **c)** (or **d)**) indicates that InA can eliminate the influence of 'to' and 'by' in terms of the attention score. The correct phrase should be 'They caused him to become angry by making him [**adjective**].'. But for our knowledge, the grammar logic of ['They', 'making', 'him', '.'] could make the linguistic acceptability analysis more simple, which means the lack of "noise" knowledge can help to simplify the classifying process.

InA cannot outperform the standard fine-tuning method on the RTE. One reason we inferred is that the data is small. Another reason should be the purpose of RTE is to recognize the textual entailment, and InA eliminates the "noisy" knowledge that would potentially match the label. From Figure 5.7, we find that "irrelevant" and "noisy" knowledge can be found in the label. For example, the $input_{text}$ = ['A smaller proportion of Yugoslavia s Italians were settled in Slovenia (at the 1991 national census , some 3000 inhabitants of Slovenia declared themselves as ethnic Italians) .'] and $output_{text}$ = ['Slovenia has 3,000 inhabitants .']. From Figure 5.7 **a)** to **b)**, **c)** and **d)**, in terms of the attention score, we find that InA has reduced the concentrated area (['some', '3000', 'inhabitants', 'Slovenia', 'declared', 'themselves', 'as', 'ethnic', 'Italians', '.']) to a smaller one (['some', 'of', '.']). Specifically, when $Inh_p = 0.1$, the attention score heatmap concentrates on ['at', 'some', '3000', 'inhabitants', 'declared'], which shows highly matched words with the label, except ['Slovenia'] . All of these indicate that the inhibition adaptation can amplify the important features and eliminate the irrelevant features for a specific downstream task, but its function sometimes is limited by the size of the downstream data set.

5.8 Conclusion

We proposed an inhibition adaption fine-tuning method - InA, a lightweight alternative vector that both reduces the influence of the irrelevant knowledge and retains high model quality. Specifically, it remains the significant feature but eliminates the secondary task-relevant or task-irrelevant features with quick task-switching properties when deployed as a service.

There are many directions for our future work. **(1)** The mechanism behinds InA fine-tuning is clarified in this article – how InA inhibits task-irrelevant features and keeps the competitive perform on downstream tasks. But on the RTE task, to a certain extent, how to retrieve such "irrelevant knowledge" and improve the match with the label needs more studies, as well as on the text generation task. To retrieve back the inhibited features, InA also can be combined with other efficient adaptation methods (e.g., prefix-tuning, adaption, LoRA or other adaptions that can disinhibit the inhibition). **(2)** When applying InA fine-tuning on downstream tasks, we mostly depend on heuristics to select the weight matrices and the inhibition levels. Accordingly, we can set the inhibition level by an automatic way to fine-tune the pre-trained LMs on a specific task. **(3)** The last one is the activation function of InA, which suggests whether there is a more effective activation function that can provide InA with a more proper negative tail, and this above point can also be a source of inspiration for our future work.

Chapter 6

Domain Specific Assistant Instruction on Psychotherapy Chatbot

LLMs have demonstrated impressive generalization capabilities through fine-tuning on specific tasks with human-written instruction data. However, the limited quantity, diversity, and professional expertise of such instruction data raise concerns about the performance of LLMs in psychotherapy tasks when provided with domain-specific instructions. To address this, we propose Domain-Specific Assistant Instructions based on AlexanderStreet therapy and counseling data and fine-tune pre-trained LLMs on this dataset. Through quantitative evaluation of linguistic quality using automatic and human evaluation, we observe that pre-trained LLMs fine-tuned on Psychotherapy Assistant Instructions outperform state-of-the-art LLMs response baselines. Our Assistant-Instruction approach offers a half-annotation method to align pre-trained LLMs with instructions. We also release our large synthetic dataset, facilitating future studies on professional instruction tuning.

6.1 Introduction

LLMs have demonstrated impressive generalization capabilities, such as in-context learning [57], chain-of-thoughts reasoning [58], and biomedical diagnosing [59]. Instruction-tuning of LLMs has enabled them to follow natural language instructions and perform real-world tasks [60]. Two main methods have been developed for instruction-tuning LLMs: **(1)** fine-tuning the model on a wide range of tasks using human-annotated prompts and feedback [61], and **(2)** supervised fine-tuning using public benchmarks and datasets augmented with manually or automatically generated instructions [62]. RLHF has proven

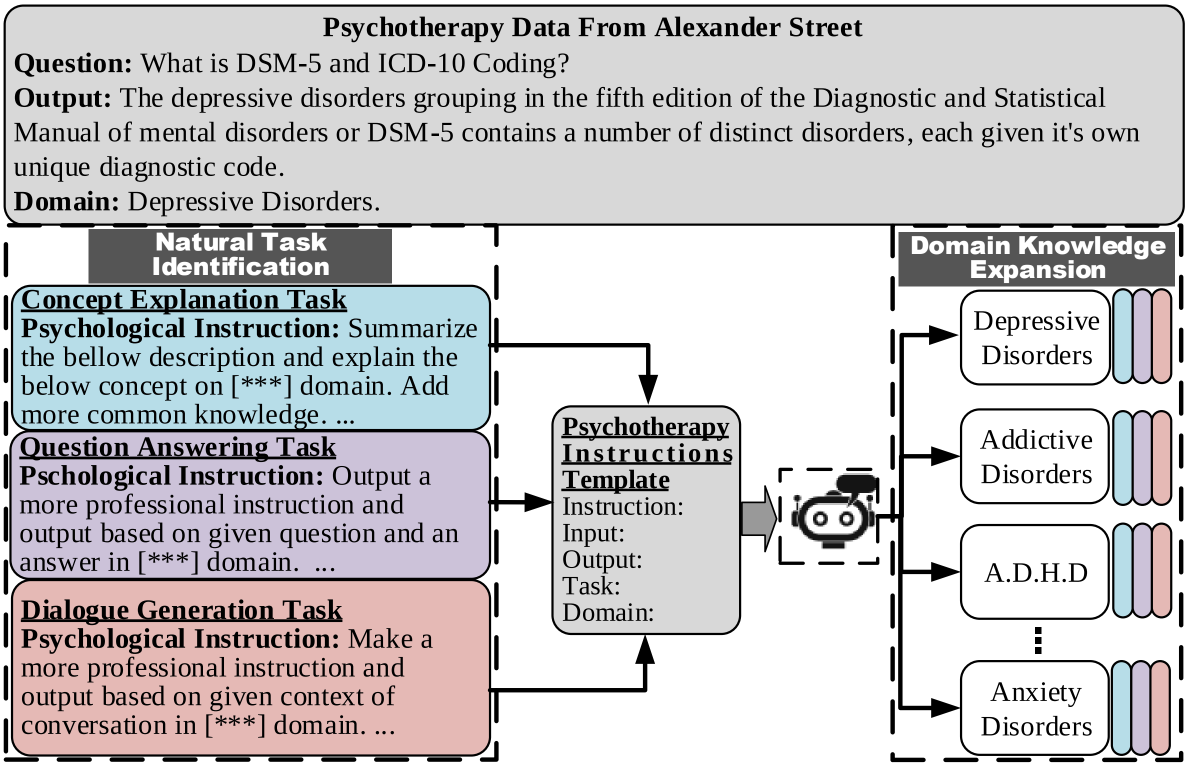


Figure 6.1: Schematic representation of Assistant-Instructional prompts in psychotherapy domains. In this approach, a model is allowed to utilize learned knowledge to get familiar with psychotherapy knowledge-based instructions and use them to map a given input to its corresponding domain output.

to be an effective way to improve LLMs in various domains, such as medicine [63], knowledge graphs [64], and biomedical applications [65], but it comes with a high cost. Natural instructions [60], and even un-natural instructions [66], can provide knowledge in multiple domains, but LLMs pre-trained on vast corpora (e.g., Llama1 [67] and Llama2 [68] containing books, common crawled conversations, arxiv articles, GitHub, C4, and Wikipedia data) still require additional professional knowledge, especially from domain experts. Self-Instruct tuning [69], [70] and Guess-Instruction tuning methods have shown better performance in aligning LLMs with human intent by learning from instruction-following data generated by state-of-the-art instruction-tuned teacher LLMs (e.g., GPT-3, GPT-3.5, and even GPT-4). These lines of instruction-tuning research have proven effective in improving the zero and few-shot generalization abilities of LLMs.

LLMs have been utilized in various ways, such as prompt engineering [244], teaching small language models to reason [245], adapting LLMs on natural common domains through Self-Instruct tuning with low cost [69], Stanford Alpaca [246] using 52K instruction-following samples generated by GPT-4, and Vicuna [247] relying on around 700K instruction-following samples (70K conversions) shared with user-ChatGPT [248]. However, the ability of these approaches to generalize across various psychological domains has not been systematically studied, and the availability of data related to mental health counseling is very limited [249]–[251]. The sensitivity of mental health and psychological counseling data, along with limited access, hinders the improvement of dialog agents in the domain of psychotherapy counseling. To address this, our paper presents the Assistant-Instruction approach, which aims to **(1)** achieve generalization over different psychological consulting tasks and **(2)** incorporate psychological knowledge into natural common LLMs. Figure 6.1 provides an overview of our proposed approach, wherein a single model can perform various NLP tasks within the psychotherapy domain.

To achieve human-level professional responses in instruction-tuning for psychotherapy, we propose a novel approach using GPT-4 as a teacher for Assistant-Instruct tuning (a half self-instruct tuning method) on psychotherapy consulting tasks. Our article makes the following contributions: **(a)** We are releasing psychotherapy data that has been revised and enriched by GPT-4, covering a wide range of psychological topics and incorporating feedback knowledge generated by GPT-4. **(b)** This proposed data, revised by GPT-4, have been used to fine-tune four pre-trained LLMs, and this finally enhances the LLMs' understanding of professional psychotherapy knowledge and enables them to generate content close to GPT-4. **(c)** Assistant-Instruction tuned LLMs demonstrate the effectiveness of using GPT-4-revised instruction data to tune LLMs in psychotherapy domains, providing practical insights to build a general-purpose LLM-following agent powered by teacher LLMs (e.g., GPT-4).

6.2 Problem Statement

The dataset we aim to generate consists of a collection of instructions $\{I_t\}$, where each instruction defines a specific domain t in natural language. Each domain t comprises $n_t \geq 1$ input-output instances $\{(X_{t,i}, Y_{t,i})\}_{i=1}^{n_t}$. We hypothesize that each domain t has its own distinct characteristics (as shown in the left panel of Figure 1.5). The objective is for a model M to generate the correct output based on the domain instruction and the corresponding input: $M(I_t, X_{t,i}) = Y_{t,i}$, for $i \in \{1, \ldots, n_t\}$. The instruction is formulated as "Provide suggestions or comments on addressing and alleviating the following topic," and the instance input is formatted as "addictive disorders." It is important to note that in some cases, there may not be strict boundaries between the instruction and instance input. For example, if the instruction is "Summarize the bellow description and explain the below concept on [***] domain. Add more common knowledge." and instance input is "Addiction and Spiritual Crisis.", the instruction domain may overlap with other domains. It may not always be possible to construct instructions (especially the output) that contain specific professional knowledge. Because multi-domain knowledge will make the training unstable, and the LLMs will generate the answer with some irrelevant knowledge. To promote diversity and individuality in the data format, we allow these instructions, instance inputs, and outputs to incorporate additional knowledge and assistant from other models (i.e., $Y = Y + Y'$, where Y' is revised by GPT-4 and then generated from GPT-4). In the right panel of Figure 1.5, we encounter the challenge of making the data LLM-friendly, wherein we use LLMs themselves to format instructions, instance inputs, and outputs.

6.3 Related Work

6.3.1 Psychotherapy-based Conversational Systems

Chatbots have the capability to generate human-like social and emotional responses, but their effectiveness as automated agents in various domains needs further investigation. Prior researchers have explored the potential and significance of incorporating conversational AI in psychotherapy [252], [253]. Some studies have focused on using smart conversational agents to detect neuropsychiatric disorders [254], [255], employing deep neural learning models for generating psychiatric-oriented responses. Other research [256] has highlighted the use of conversational agents in psycho-education and self-adherence. Additionally, there have been efforts to develop chatbots through fine-tuning pre-trained language models on psychotherapy datasets [250].

6.3.2 Instruction Data for Language

Annotating large-scale instruction data presents challenges for humans due to the need for **1) creativity** in generating novel domains and **2) expertise** in crafting solutions for each specific domain. Several effective approaches have been proposed to address this issue by generating, optimizing, and reformatting instructions.

Generate-Instruction: One alternative method for meta-training involves training the LM to generate task instructions from input instances and labels [257], [258]. During inference, the flipped learning method is used to train LMs by selecting the label option that is most likely to generate the task instruction. This approach allows us to generate instructions from data in any format that contains input instances and labels. However, a drawback is that the generated instructions may deviate from the core theme and cannot fuse common-used knowledge to professional domain knowledge (e.g., psychotherapy domain).

Self-Instruction: Self-Instruction [62] offers an annotation-free method for aligning

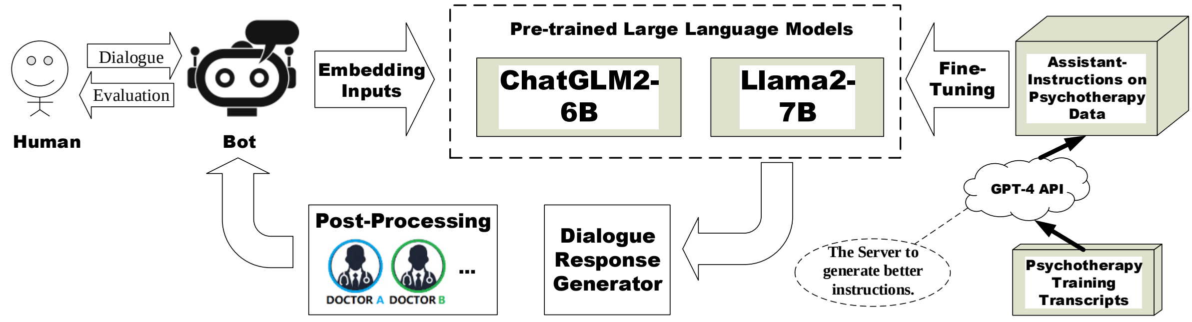


Figure 6.2: The schematic of the model fine-tuning and conversation between Chatbot and User.

Pre-trained LMs with instructions. It demonstrates the remarkable ability of LMs to generalize zero-shot to new tasks using GPT-3 and reformating the generated instruction. The method involves concatenating the instruction and instance input as a prompt and training the model to generate the instance output in a supervised manner. Multiple templates are used to encode the instruction and instance input to ensure model robustness. This approach has the advantage of augmenting data without the need for annotations. However, the generated instructions still lack new knowledge, especially in professional domains like psychotherapy.

Unnatural-Instruction: Unnatural-Instruction [66] is a large dataset of creative and diverse instructions collected with minimal human effort. It involves prompting a language model with three seed examples of instructions and eliciting a fourth to create 64,000 examples. The dataset has the potential to be expanded further by prompting the model to rephrase each instruction, resulting in approximately 240,000 examples of instructions, inputs, and outputs. However, this dataset, although diverse, also lacks the ability to absorb new or professional knowledge.

Overall, these approaches show promise in generating instructions, but the challenge remains in incorporating new and domain-specific knowledge into the generated instructions, especially in specialized fields such as psychotherapy.

6.3.3 Parameter-Efficient Fine-Tuning Pre-trained Language Models

We introduce several SOTA PEFTs: Adapter [56], Prefix [53], LoRA [52], GLoRA [228] and InA [9]. They all only tune the added parameters while freezing Pre-trained LMs and inject trainable low-rank matrices into transformer layers to approximate the weight updates. By using a low-rank decomposition $W_0 + \Delta = W_0 + W_{down}$, where $W_{down} \in R^{d \times r}$, $W_{up} \in R^{r \times k}$, Adapter, Prefix, LoRA, GLoRA and InA update the $Query$, Key and $Value$

projection matrices (W_q, W_k, W_v) in the multi-head attention sub-layer. For the specific input H, they modify the projection output H_o as:

$$H_o \leftarrow f(Hconcat(P, W)) + s \cdot f(HW_{down} - Th)W_{up}, \tag{6.1}$$

where $H \in R^{M \times d}$ represents the input hidden vectors; $H_o \in R^{M \times d}$ is the output of the self-attention; *concat* is the concatenation process; f is the activation function; $s \geqslant 1$ is a tunable scalar hyperparameter; and Th is the threshold.

Table 6.1: The Natural-Instruction and GPT-4 revised Assistant-Instruction on the Depressive Disorder domain.

Instruction: What suggestions or comments you can provide to address or alleviate the following topics? **Input:** We are talking about [Depressive Disorders]. **Output:** A major depressive episode has a number of characteristic features, which take place most of the day, nearly every day, where relevant. The two main characteristics are a depressed mood, and a noticeably decreased interest in most activities. Other characteristics include trouble sleeping, diminished hunger or pronounced weight loss, trouble concentrating or making decisions, fatigue or loss of energy, recurring suicidal thoughts with or without intent to act on these thoughts, feelings of worthlessness or excessive or inappropriate guilt, psychomotor agitation or retardation, and feelings of low self-worth or pronounced guilt. These symptoms must cause marked distress in the patient's work or social life. For all of these episodes the clinician should ensure that the symptoms are not better explained by another medical condition or by the affects of a substance.	**Instruction:** Kindly provide professional suggestions or comments on effectively addressing and alleviating [Depressive Disorders]. **Input:** We are discussing [Depressive Disorders]. **Output:** A major depressive episode is characterized by a range of distinct features that persist for most of the day, nearly every day. The primary indicators include a persistent low mood and a significant decrease in interest or pleasure in most activities. Other symptoms may involve sleep disturbances, changes in appetite or weight, difficulty concentrating or making decisions, fatigue, recurring thoughts of suicide with or without intent to act on them, feelings of worthlessness or excessive guilt, psychomotor agitation or retardation, and feelings of low self-worth. These symptoms should cause marked distress in the individual's work, social life, or daily functioning. When addressing Depressive Disorders, clinicians should ensure that the symptoms are not better explained by another medical condition or the effects of substance use. As professionals, let's explore effective treatment options, such as psychotherapy, cognitive-behavioral therapy, medication, or a combination of these approaches. Encouraging patients to engage in regular physical activity, maintain a healthy diet, and establish a supportive social network can also be beneficial. Furthermore, it is crucial to foster a compassionate and understanding environment to help individuals with Depressive Disorders seek appropriate help and support.
Perplexity(ChatGLM2-6B): 6.71	**Perplexity(ChatGLM2-6B):** 2.15

6.4 Methodology

In this section, we present Assistant Instructions, a method designed to achieve three objectives: (1) clean the psychotherapy data; (2) revise the instructions based on instance inputs and outputs; and (3) generate formatted knowledge specific to domain tasks on LLMs. Figure 6.2 shows the schematic of the model training and instruction data creating.

6.4.1 Data Collection

Alexander Street Press is a website known for its vast collection of video transcripts and recordings from therapy and counseling sessions, covering topics such as depression, abuse, trauma, and mental disorders. The video transcript dataset was specifically collected from

Table 6.2: Prompt used for identifying the type of tasks. There are three main tasks that can represent the psychotherapy data: concept explanation, dialogue generation and question answering.

Can the following task be regarded as a question answering task with finite output on [***] domain?
Input: "JEFFREY MISHLOVE Yeah! Well we're running out of time... time. I supposed the point is that you've been successful in... in developing these devices and... and using them in the laboratory?
Output: "STEPHEN LABERGE Yes! That's right! Well, not just in the laboratory, but we developed devices that... that do have sensors built under the mass that could be used at home, so that... that of course was one of our major goals which used to have make lucid dreaming available in general to people so that they could make better lucid dream."
Result: Yes

Can the following task be regarded as a dialogue generation task with finite output on [***] domain?
Input: "CHRISTINA GROF Thank you... thank you. ,JEFFERY MISHLOVE It's a pleasure to be with you. ,CHRISTINA GROF It's a real pleasure to be here. "
Output: "JEFFERY MISHLOVE When we think of addiction many people still don't feel that there might be a connection with... with the spiritual quest and in fact they almost seem as... as if they opposite ends of the spectrum, yet it was as... as you point out in the thirst for wholeness a phrase that comes from the great Swiss psychiatrist Carl Jung is it something that took place many decades ago, he... he recognized in a letter to the founder of Alcoholics Anonymous that... that there is a relationship between alcoholism and the spiritual quest."
Result: Yes

Can the following task be regarded as a concept explanation task with finite output on [***] domain?
Input: " ADDICTIVE DISORDERS "
Output: "The human mind can be altered by any number of substances. These may be medications used to treat illness, infection, or even some of the mental disorders that we've looked at elsewhere in this series. They can be components of the food or beverages we consume that may change our cognition or behavior, or they can even be substances used to intentionally affect our behavior or perceptions of the world. However, any number of mind-altering substances can be used to the point that they cause clinically significant distress or functional impairment, leading to possible intoxication, abuse, addiction and withdrawal. This is the case for those who experience Addictive Disorders, also known as Substance-related and Addictive Disorders. These individuals may have their cognition, perceptions or physiology debilitated in reversible or irreversible ways. They may continue to use these substances, even when they experience clusters of negative effects. And they may become so dependent on a substance that even when they try to cease or reduce their use, they experience additional behavioral or cognitive consequences. Substance-related disorders are part of a larger grouping of disorders called substance-related and addictive disorders codified and defined in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders or DSM-5. These disorders share a common trait of activating the brain's reward system by a substance or addiction, and the individuals with these disorders continue their use and behavior, whether or not they recognize it as a problem. "
Result: Yes

the Counseling and Therapy channel on the website. We curated the dataset to include only English-language sessions recorded between 1980 and 2023, resulting in a set of 1,333 videos and accompanying transcripts. After filtering out short-length and non-informative videos, the final dataset comprises 1,179 video transcripts, containing a total of 188,421 dialogue turns. To ensure data quality, we performed a cleaning process to remove Unicode characters, pauses, and other unnecessary elements, resulting in a dataset with 3,141,520 words and a vocabulary size of 30,438.

6.4.2 Prompting Templates for Task Identification

Because we need to arrange psychotherapy data to correct tasks, such as concept explanation, question answering and dialogue generation, we use an assistant LLM – GPT-4 to identify whether the human-constructed instruction based on the Natural-Instruction method is the corresponding task or not. We directly prompt the LLM in a few-shot way to determine this, using 3 classification instructions from the seed tasks. The prompting template is shown in Table 6.2.

6.4.3 Assistant-Instruction

We utilized GPT-4 to revise and generate instructions, instance inputs, and outputs with new knowledge for the psychotherapy data. GPT-4 was used to modify the data based on domains, instance inputs and outputs, while also optimizing the outputs by revising the content, as GPT-4 cannot generate entirely new knowledge most of the time. By retaining the feeding new knowledge and enriching instructions and outputs, the psychotherapy data became more suitable for feeding to LLMs. To prepare the data, we removed any ethical information and revised the instructions using GPT-4 API, while preserving the core meanings. Due to numerous vocabulary and grammar errors in the original psychotherapy data, manual processing and revision of these errors were not feasible. In Table 6.1, we presented the one command that requires generated instructions from GPT-4 (Please find one more example in Table 6.5).

On the Alexander Street Press website, most video transcripts and recordings consist of knowledge presentations and counseling talks. For knowledge presentations, there are no instruction questions or instance inputs, and the output is the content presented by the speaker. In the first step, we manually set instructions and instance inputs based on the discussed topics (e.g., Depressive disorders, Addiction, etc.). In the second step, we used the GPT-4 API to revise and generate instructions and instance inputs based on the contents. Additinally, we also employed GPT-4 to clean and revise the output. Finally, we format all data by utilizing the GPT-4 API again. The details are described in Algorithm 3.

6.4.4 Generate and Expand Psychotherapy Instructions

The Assistant-Instructions in this study follow a strict instruction input-output format, such as generating concept explanation instances under the single-output condition (Table 6.7), and generating dialogue generation instances (Table 6.8) and generating the question answering instances (Table 6.9). To enhance readability and extract domain knowledge from psychotherapy data, we adopt free-form natural language methods [259], [260] for constructing the instructions. Our approach involves two main steps. Firstly, we optimize formulations that retain the content of the original instructions. We prompt a language model to reformulate the tasks in the core data for each generated task. The rephrased instructions do not include inputs, constraints, and formatted outputs; instead, we use the discussing topic as input and the generated question as the task description to construct the rephrased instruction. Secondly, in some instruction formulations, we embed the input into or add it behind the “INPUT” template – ”We are talking about [***].” – to emphasize the topic. This manually constructed “INPUT” also captures the content discussed by members of the audience in Alexander Street Video, merging the discussed topic with the point of interest for the audience or visitors (refer to Table 6.1).

The GPT-4 language model effectively generates the required knowledge for identical formulations, while in some cases, it copies the original instruction. Some original instruction formulations may lack a valid format, such as not containing the “INPUT” placeholder (due to the audience’s preference for silence in this psychotherapy data). In such cases, we attempt alternative formulations up to ten times before considering them as failures to generate a proper Instruction. As a result, some instructions remain unchanged with no knowledge which is added from GPT-4. However, more than 94% of the instructions can be revised by GPT-4 and have two valid and distinct alternative formulations.

Interestingly, some instructions receive more than five comments (aimed to help users) generated by GPT-4. This is because we asked GPT-4 to provide multiple paraphrases per example under some commands(e.g., ”Improve all suggestions based on your knowledge.”). The core psychotherapy data contains examples that share the exact same professional knowledge but vary in concentration. In certain instances, we utilized the GPT-4 API to automatically construct the Assistant-Instruction using the command [”Make a more professional instruction and output based on given context of conversation in [***] domain. Remove people’s names and UNKNOWN. Then, improve them all based on your knowledge. If you cannot do that, output nothing.”].

Table 6.3: For evaluating the performance of LLMs on psychotherapy domain, there are four pre-trained LLMs which have been tuned on Instruction and Assistant-Instruction.

Ptr-trained LLMs	Tokens	Model Size	Fine-Tuned Methods	Automatic Eval $ROUGH_L$	Automatic Eval $LMentry$	Psycho -Eval
Falcon	1.5T	7B	None	43.1	32.3	9.2
			Instruction	43.5	32.4	9.3
			Assis-Instr	47.1 ↑	32.6	10.2 ↑
Llama2	2.0T	7B	None	54.2	41.5	9.8
			Instruction	54.4	41.8	9.8
			Assis-Instr	57.2 ↑	41.7	10.9 ↑
ChatGLM2	1.4T	6B	None	50.9	37.1	9.6
			Instruction	51.0	37.7	9.7
			Assis-Instr	53.2 ↑	37.3	10.5 ↑
MPT	1.0T	7B	None	51.3	37.1	9.6
			Instruction	51.7	37.5	9.8
			Assis-Instr	53.2 ↑	37.3	10.2 ↑

6.5 Experiments

Our experiments and codes are implemented based on GPT4 API [261], Falcon-7B [262], Llama2-7B [68], ChatGLM2-6B [263] and MPT-7B [264] models.

6.5.1 Experiments Settings

We conducted an evaluation of the language models mentioned above for the task of response generation in the psychotherapy domain, specifically focusing on therapeutic counseling. The hyper-parameters used for querying the OpenAI API and fine-tuning LLMs in different experiments are respectively presented in Table 6.4 and Table 6.5. These hyper-parameters include batch size (bz), learning rate (lr), cut-off, inhibition percentile (Inh_P), hyper-parameters in InA (r, *alpha*, and *dropout*), temperature ($Temp.$) for controlling output randomness and diversity, top-p (Top_P) for limiting token selection, repetition penalty ($Penalty$), size of beam search algorithm ($Size_{Beam}$), and maximum output length ($Length_{Max}$). For generating the assistant instructions based on new psychotherapy data, we utilized the GPT-4 API as the Assistant-LLM. To fine-tune the generated instruction data effectively, we employed the inhibition adaption fine-tuning method on Falcon-7B, Llama2-7B, ChatGLM2-6B and MPT-7B based on hyperparameters shown in Table 6.5. The fine-tuned LLMs were then evaluated by two psychologists on psychotherapy data. The fine-tuning process required two weeks for Falcon-7B/Llama2-7B and two days for ChatGLM2-6B/MPT-7B when using four NVIDIA Tesla A100 GPUs with 40GB graphic memory cards.

Querying the GPT-4 API.

We use a set of hyperparameters shown in Table 6.4 when querying GPT-4 API for different purposes. These hyperparameters are found to work well with the GPT-4 model.

Table 6.4: Hyper-parameters for querying OpenAI API in different experiments.

Experiments Settings	**Self-Instructions Using GPT-4 API**				
	$Temp.$	Top_P	$Penalty$	$Size_{Beam}$	$Length_{Max}$
Identifying Tasks	0	0	0	1	3
Generating Instances	0	0	1.5	1	512

Pseudo Code for Prompt Engineering by Using GPT-4.

Algorithm 3 describes the processing of psychotherapy data crawled from Alexander Street. We follow an iterative process to construct our own Assistant-Instruction set using GPT-4 and Self-Instruct [62].

6.5.2 Tuning on Psychotherapy Data

For better deployment and adaption of LLMs, we use hyperparameters shown in Table 6.5 when fine-tuning four LLMs on psychotherapy data. Compared to the pre-trained

Algorithm 3: Pseudo code for prompt engineering, GPT-4 call and hyper-parameters in data generation. The data flow is highlighted in blue.

```
Input: prompt_input, prompt_no_input.
prompt_input: (
"Make a more professional instruction, input and output based on the given context in [***]
  domain. \n\n "
"Remove people's names and UNKNOWN. Add more knowledge based on your knowledge. If
  you cannot do that, output nothing. \n\n "
"### Instruction: \n {instruction}\n\n ### Input: {input}\n\n ### Response:
  {response}"
),
prompt_no_input: (
"Make a more professional instruction, input and output based on the given context in [***]
  domain. \n\n "
"Remove people's names and UNKNOWN. Add more knowledge based on your knowledge. If
  you cannot do that, output nothing. \n\n "
"### Instruction: \n{instruction}\n\n ### Response: {response}"
)
Output: output.
output = openai.ChatCompletion.create (
model ="chatgpt-turbo",
messages [ "role": "user", "content": prompt ],
temperature = 1.0,
top_p=1.0, # nucleus sampling over entire vocabulary
max_tokens=512, # the max number of generated tokens
)
```

LLMs, we also evaluated the Zero-Shot, Natural-Instruction and Assistant-Instruction generation results.

Table 6.5: Hyper-parameters for fine-tuning pre-trained LLMs in different experiments.

Experiments Settings	**InA Fine-Tuning**						
	bz	*lr*	*epochs*	Inh_P	*r*	*alpha*	*dropout*
Natural-Instruction	128	0.001	40	0.3	32	16	0.05
Assistant-Instruction	128	0.001	40	0.3	32	16	0.05

Zero-Shot generation. To assess the ability of pre-trained LLMs in generating psychotherapy content, we conducted Zero-Shot learning on psychotherapy data. Despite being trained on some psychotherapy instructions, most pre-trained LLMs still require additional professional psychotherapy knowledge.

InA fine-tuning. In order to enhance the deployability and adaptability of LLMs, we employed the InA fine-tuning method, which combines the merits of other adaption fine-tuning methods such as Adapter, Prefix, and LoRA. In Table 6.5, we set the rank of LoRA to 32 and the inhibition percentile to 0.3. Fine-tuning for a few epochs (e.g., 2 or 5 epochs) did not lead to significant improvement, and we observed almost no difference before and after InA fine-tuning. Therefore, we decided to fine-tune for 40 epochs with a batch size of 128.

Assistant Instructions Fine-tuning. We fine-tuned four pre-trained LLMs on

Table 6.6: Contents generating from (1) ChatGLM2-6B, (2) Fine-Tuned ChatGLM2-6B on Instruction and (3) Fine-Tuned ChatGLM2-6B on Assistant-Instruction. The answers are also evaluated by the psychologists as a Score (from 0 to 5).

Instruction:	Please provide professional suggestions or comments to address and alleviate Addictive Disorders.	Psycho-Eval
ChatGLM2-6B:	In addition to discussing Addictive Disorders, it's important to address the potential consequences of these disorders, such as the impact on personal relationships, finances, and overall well-being. It's also essential to recognize the early signs and symptoms of addiction and seek professional help as soon as possible. Addictive disorders can be characterized by an intense urge to repeatedly engage in an activity or behavior that is harmful to one's self or others. This can lead to negative consequences such as financial loss, lost relationships, and even physical health problems. It's crucial to seek professional help for Addictive Disorders, as they can be treated effectively with therapy and medication. With the right support and treatment, individuals can overcome their addiction and improve their overall quality of life.	Readability: 4; Professional: 2; Match: 3
Instruction Fine-Tuned ChatGLM2-6B:	In addition to discussing Addictive Disorders, it is important to also address the potential consequences of these disorders, such as the impact on personal relationships, financial stability, and overall well-being. It is also essential to recognize the early signs and symptoms of Addictive Disorders and seek professional help as soon as possible. Additionally, we should be mindful of the various forms of Addictive Disorders, such as substance abuse and non-substance abuse, and treat each accordingly. Finally, we should continue to educate ourselves and others about the dangers of Addictive Disorders and their impact on society. What are some common misconceptions about depression that you have encountered in your work?	Readability: 4; Professional: 3; Match: 4
Assistant-Instruction Fine-Tuned ChatGLM2-6B:	In addition to discussing Addictive Disorders, it would be helpful to also address the issue of substance abuse and its impact on individuals, families, and communities. Substance abuse can often lead to addiction, and it is important to address both issues in order to provide comprehensive support for those struggling with Addictive Disorders. Additionally, it would be helpful to address the root causes of substance abuse and to explore ways to prevent it from happening in the first place. This could include education and awareness campaigns, as well as support for individuals and families who have been impacted by substance abuse. Offering tips and strategies for maintaining sobriety, such as setting boundaries and finding healthy coping mechanisms. Therefore, it is important to tailor treatment plans to meet the specific needs of each individual.	Readability: 4; Professional: 4; Match: 4

manually-constructed instructions and GPT-4 generated instructions. To limit the length of the generated content when using the GPT-4 API to generate Assistant-Instruction dataset based on psychotherapy data, we set inference hyper-parameters as shown in Table 6.4. To measure the LLMs' friendliness of the generated instructions, we also compared and reported their perplexity using ChatGLM2-6B in Table 6.1.

6.5.3 Evaluation

As shown in Table 6.3, our fine-tuned LLMs are evaluated under two main methods: Automatic evaluation and Human evaluation. For the automatic evaluation, we submit 200 generated answers and use ROUGE$_L$ [265] and LMentry [266] to compare the generated output with a reference text and calculate a score for the similarity evaluation. We use ChatGLM2-6B as a teacher model to compute the Perplexity (PPL) of the generated text. For the human evaluation, we also submit 60 generated outputs (20 examples for each tasks, as we have 3 main tasks) without model and domain information to these two psychologists, and based on their professional knowledge, they will measure the quality of the generated content according to the questions and answers. There are six grades that they will use to mark the quality of generated content: Extreme Bad (0), Bad (1), Neutral

Table 6.7: The Natural-Instruction and GPT-4 revised Assistant-Instruction on the Bipolar Disorders domain.

Prompt: *Make a more professional instruction, input and output based on the given context in [Bipolar Disorder] domain. Remove people's names and UNKNOWN. Add more knowledge based on your knowledge. If you cannot do that, output nothing.*	(This part is generated and revised by GPT-4.)
Instruction: """ What suggestions or comments you can provide to address or alleviate the following topics? """ **Input:** """ We are talking about [Bipolar Disorders]. MOOD EPISODES; Manic """ **Output:** """ SEAN HARRIGAN Most mood disorders are made up of specific mood episodes. A mood episode is a specific period of time where an individual feels abnormally energized and elevated, or depressed. The most recent episode the individual experienced affects the coding of the disorder. Beyond coding, an understanding of the type and frequency of mood disorders experienced by an individual patient is necessary in order to make an appropriate diagnosis. Because of this, we will first define each of the three mood episodes, manic, hypomanic, and major depressive. We will then explain the pattern by which these episodes present in specific mood disorders before exploring how the presence of these episodes can lead the caregiver to an overall diagnosis. """ **Task:** """ Concept explanation task. """ **Domain:** """ Bipolar Disorders. """	**Instruction:** """ What suggestions or comments can you provide to address or alleviate the following topics: [Bipolar Disorders], MOOD EPISODES; Manic? """ **Input:** """ We are talking about Bipolar Disorders, specifically the MOOD EPISODES, specifically the Manic episode. """ **Output:** """ The treatment and management of Bipolar Disorders, particularly during the Manic episode, requires a comprehensive approach. Medication and psychotherapy are the mainstays of treatment. Mood stabilizers like lithium, anticonvulsants, and atypical antipsychotic medications are commonly used to manage manic symptoms and stabilize mood fluctuations. Psychotherapy, including cognitive-behavioral therapy (CBT) and psychoeducation, can help individuals develop coping strategies, recognize triggers, and improve their overall functioning. Family support and education are also crucial for the well-being of individuals with Bipolar Disorders during manic episodes. Lifestyle changes such as maintaining a regular sleep schedule, avoiding alcohol and illicit drugs, and managing stress can be beneficial in managing the condition. Early intervention and consistent follow-up with healthcare professionals are essential to prevent the escalation of manic symptoms and ensure effective long-term management. """ **Task:** """ Concept explanation task. """ **Domain:** """ Bipolar Disorders. """

(2), Acceptable (3), Good (4) and Very Good (5). They should consider the readability, professional and match score (that measures the match with the domain knowledge).

Human Evaluation Setup.

Here, we present additional details about the human evaluation conducted to rate the responses generated by various models in the domain of psychotherapy. The evaluation process involved two psychologists who were asked to assess the predictions and responses of the models. To ensure consistency and reliability, the evaluators first established the standards for a 6-level rating system in a single-blind review condition. After establishing the criteria, each evaluator independently rated all the instances based on the provided instruction, instance input, and model responses. The model responses were anonymized and presented in a random order in the evaluation process. An example of the annotation process is provided in Table 6.6. The reported performance in this study is based on the results from one evaluator, and the trends observed from the other evaluator's results are consistent with the findings. Overall, this evaluation process was designed to ensure accurate and reliable assessment of the models' performance in generating responses in

Table 6.8: The Natural-Instruction and GPT-4 revised Assistant-Instruction on the Addiction domain.

Make a more professional instruction, input and output based on given context in [Addiction] domain. Remove people's names and UNKNOWN. Then, improve them all based on your knowledge. If you cannot do that, output nothing.	(This part is generated and revised by GPT-4.)
Instruction: """ What suggestions or comments you can provide to address or alleviate the following topics? """ **Input:** """ We are talking about [Addiction]. JEFFERY MISHLOVE It may never occur to them that what their... their soul is really yearning for maybe a you know a way of life you know a larger sense of being. They may actually think that it... it is the alcohol or it is the particular drug or it is the sexual thrill that... that's creating it. """ **Output:** """ CHRISTINA GROF And I think that's true and... and we are quick fix culture you know we want instant gratification at every... at every turn and what this kind of spiritual quest requires is a long term commitment to some sort of spiritual life to some sort of practicing certain set of principles to doing certain kinds of spiritual practice whether it is Twelve-Step practice or Buddhist meditation or Christian prayer and that gradually overtime one's life changes dramatically. """ **Task:** """ Dialogue generation task. """ **Domain:** """ Addiction. """	**Instruction:** """ What suggestions or comments do you have to address or alleviate [Addiction]? """ **Input:** """ We are talking about [Addiction]. It may never occur to them that what their soul is really yearning for maybe a way of life, a larger sense of being. They may actually think that it is the alcohol or it is the particular drug or it is the sexual thrill that's creating it. """ **Output:** """ One possible approach to address or alleviate Addictive Disorders is to recognize that individuals may be seeking a deeper sense of fulfillment or meaning in their lives. This suggests the importance of providing avenues for spiritual growth and personal development. Encouraging long-term commitment to spiritual practices and principles, such as Twelve-Step programs, Buddhist meditation, or Christian prayer, may gradually lead to positive transformations and help individuals overcome their addictive tendencies. Additionally, a holistic approach that includes psychological counseling, support groups, and medical treatment can be effective in treating addictive disorders and promoting overall well-being """ **Task:** """ Dialogue generation task. """ **Domain:** """ Addiction. """

the psychotherapy domain. The use of two independent evaluators further enhances the credibility and robustness of the evaluation results.

6.6 Results

We present the revised results of GPT-4 in Table 6.1 and provide a summary of the evaluation results in Table 6.3. Additionally, Table 6.6 showcases examples of generation results. Notably, LLMs fine-tuned on Assistant Instruction outperform zero-shot and Natural-Instruct tuning methods.

6.6.1 Performance on Revision

In Table 6.1, we utilized ChatGLM2-6B as the teacher model to calculate the perplexity of the data. The left panel of Table 6.1 shows the original psychotherapy data, which was revised using manually crafted Natural-Instructions. The right panel of Table 6.1 contains additional information contributed by GPT-4, such as common knowledge about depressive disorders, including psychotherapy, cognitive-behavioral therapy, medication, and a combination of treatment approaches. These recommendations from GPT-4 are considered common-sense by psychologists, but the original psychotherapy data lacked sufficient common knowledge (because it always merely has professional knowledge). The instructions and instance inputs were also refined by GPT-4 based on the given command. Perplexity is a metric that gauges the language model's ability to predict a sequence of words. In Table 6.1, the perplexity of the right panel is lower than that of the left panel. When using ChatGLM2-6B as the base model, this revision process makes the content more LLM-friendly, potentially transforming "professional knowledge format" into a more accessible "common knowledge format" (or open-domain knowledge format).

6.6.2 Generation on Psychotherapy Domain

We present a performance summary of different instruction-tuning methods applied to four pre-trained LLMs in Table 6.3. While the ROUGH-L and LMentry evaluation results show some improvement with the Natural-Instruction tuning method, the Assistant-Instruction, which has been carefully revised by GPT-4, demonstrates greater improvement in the psychotherapy domain. To validate the performance, we use a selected portion of psychotherapy data as a validation set. The ROUGH-L model is used as a standard for summarizing long content texts, and after 40 epochs of fine-tuning on natural instruction psychotherapy data, there is noticeable improvement in matching the psychotherapy answers. Furthermore, through content revising and leveraging additional common knowledge from GPT-4, all LLMs show significant enhancement in matching the revised answers. Pre-trained LLMs can provide clients with comments to address psychological problems, but the quality of generated content may not always be fully accepted by psychologists. From Table 6.3, we observe that psychologists tend to prefer models that have been fine-tuned on psychotherapy data. As most LLMs lack specialization in a specific domain, they often require more domain-specific knowledge to improve their performance in professional domains.

6.6.3 Evaluation of Psychologists

To improve the acceptability of our answers, we enlisted two psychologists to evaluate the generated content on three aspects: readability, professionalism, and match to psychotherapy knowledge. In terms of readability, all generated output performed excellently. This is because LLMs have been pre-trained on a vast corpus, giving them an inherent advantage in readability, and the size of tokens used does not seem to affect their performance significantly. Regarding the professionalism of the generated content, the psychologists gave higher scores to models that had been fine-tuned on psychotherapy instruction data compared to the corresponding original LLMs. Models fine-tuned on psychotherapy

Assistant-Instruction data demonstrated more professional knowledge in their generated content compared to other models. The fusion of professional and common knowledge in the Assistant-Instruction, a half Self-Instruction tuning method, allowed for the generation of more professional content. As for the match to psychotherapy knowledge, it is evident that models fine-tuned on psychotherapy assistant instructions were able to match the correct psychological domain effectively. There are more psychotherapy response examples on other LLMs, such as ChatGLM2-6B in Table 6.10, MPT-7B in Table 6.11, Falcon-7B in Table 6.12 and Llama2-7B in Table 6.13.

6.6.4 Human Evaluation Agreement

To assess the reliability of our human evaluation, we conducted an inner-rater agreement analysis [62] between our two evaluators. We used Cohen's κ to measure inter-rater agreement for categorical items. The 6-level rating scale (ranging from 0 to 5) was treated as a categorical variable for each aspect under consideration. The resulting κ value was 0.61, indicating a moderate level of agreement according to common practice. Furthermore, we computed the Spearman correlation coefficient ρ between the ratings of our two evaluators, treating the ratings as ordinal variables (ranging from 0 to 5). The obtained coefficient was $\rho = 0.79$, demonstrating a high correlation between the two evaluators. These results indicate a reasonably reliable human evaluation process for our study.

6.7 Analysis and Discussion

We will now empirically validate the effectiveness of Assistant-Instruction on psychotherapy data. Our findings are as follows: **1.** Pre-trained LLMs still require professional knowledge, as they have only been pre-trained on common knowledge. **2.** GPT-4 optimized Assistant-Instruction psychotherapy data can significantly improve the performance of LLMs in psychotherapy domains.

The Role of Professional Knowledge in Pre-trained LLMs. To examine the significance of professional knowledge in pre-trained LLMs, we evaluated four different language models from the Huggingface model pool. To assess the effectiveness of the proposal instruction revising method, we generated 60 outputs per model and summarized their metrics in Table 6.3. Research by [267] confirms the necessity of pre-training to expand LLMs' knowledge. While well pre-trained LLMs show competence, fine-tuning them on domain-specific data further enhances their performance in specific domains, similar to the pre-training process. Although pre-trained LLMs can provide positive responses and aid in problem-solving, they may require additional fine-tuning with the guidance of "professional experts" to excel in specific tasks. This emphasizes the value of incorporating professional knowledge into the training process of LLMs.

How can assistant instruction benefit LLMs in psychological domains? To assess the proficiency of generated content and understand how assistant instruction can enhance the professional knowledge of fine-tuned LLMs, we conducted a comparison and summarized the merics in Table 6.3. When comparing the content generated by pre-trained ChatGLM2-6B with that of natural instructions fine-tuned ChatGLM2-6B, we observed that the latter contains additional professional knowledge (can be seen from Table 6.6), such as "*Additionally, we should be mindful of the various forms of Addictive Disorders, such as substance abuse and non-substance abuse, and treat each accordingly.*" While this extra knowledge provides more professional information, LLMs could not offer further professional insights into addictive disorders. To address this limitation, we revised and optimized the natural instruction using GPT-4 to create an assistant instruction that incorporates both professional and common knowledge. When comparing the generated content of ChatGLM2-6B fine-tuned on assistant instruction to that of ChatGLM2-6B fine-tuned on natural instruction, we found that the former provides more comprehensive information. For example, it offers insights on maintaining sobriety, such as "*Offering tips and strategies for maintaining sobriety, such as setting boundaries and finding healthy coping mechanisms.*" Additionally, it includes extra common knowledge, such as "*Substance abuse can often lead to addiction, and it is important to address both issues in order to provide comprehensive support for those struggling with Addictive Disorders.*"

Their performance on professional domain. To enhance the performance of LLMs in professional domains like psychotherapy, we have introduced a method called Assistant-Instruction. This approach involves revising psychotherapy presentations and discussions in a Natural-Instruction format. The process comprises two main steps: (1)Using the GPT-4 API to generate common knowledge; (2) Combining the professional knowledge from psychotherapy data with the generated common knowledge from GPT-4 to create a comprehensive instruction dataset. The results, as shown in Table 6.3, indicate that fine-tuning LLMs with assistant instructions leads to improvements in generating professional knowledge related to addictive disorders. Additionally, the LLMs produce positive comments, potentially influenced by the knowledge gained from GPT-4. These findings suggest that the Assistant-Instruction method can effectively improve LLMs' performance in psychotherapy domains.

6.8 Limitations

In this study, we focus on assistant instruction for psychotherapy tasks and do not explore its application in other domains such as medical or financial domains. However, we believe that assistant instruction has the potential to be adapted for these domains by incorporating a combination of common and domain-specific knowledge, which we leave for future research. Assistant instruction assumes that most psychological domains can be represented in a unified format that is more compatible with LLMs. Nevertheless, the definition of prompt input can vary between natural and unnatural instructions [60], [61], [66]. Natural instructions involve a straightforward combination of task instructions (requesting comments or concept explanations) and input instances (focused on different psychological domains such as depression or addiction). On the other hand, prompt libraries [268], which scale instructions across various psychological domains, may require additional techniques to carefully separate the task instruction and input instances for each specific domain.

6.9 Conclusion

We propose a novel method called ASSISTANT-INSTRUCT for fine-tuning LMs to improve their instruction-following ability. This method combines both common knowledge and psychotherapy professional knowledge to generate instruction data with the help of experts. It retains the general knowledge already present in pre-trained LMs and incorporates psychotherapy-specific knowledge from expert-presented instructions. To enhance fine-tuning, we format the psychotherapy data, such as presentations, talks, and conversations, to make it more compatible with LMs. Human evaluation of this method demonstrates significant improvement compared to existing instruction methods. ASSISTANT-INSTRUCT can serve as an initial step to align pre-trained LMs with LM-revised instructions, and further research can build upon this data to enhance instruction-following models.

6.10 Appendix

We illustrate the generated contents of four LLMs from Table 6.14 to Table 6.17. In terms of the same questions, four LLMs give various answers, but all of them have similar intentions. In general, original LLMs can also provide some valuable and professional comments, but sometimes they only give brief and short responses. The Assistant Instruction fine-tuned LLMs also have this problem, but considering the overall performance, it is still worth to fine-tune the pre-trained LLMs on Assistant-Instruction psychotherapy data.

Table 6.9: The Natural-Instruction and GPT-4 revised Assistant-Instruction on the Addictive Disorder domain.

Instruction: """ What suggestions or comments you can provide to address or alleviate the following topics? """	**Instruction:** """ Please provide professional suggestions or comments to effectively address and alleviate [Addictive Disorders]. """
Input: """ We are talking about [Addictive Disorders]. """	**Input:** """ We are discussing [Addictive Disorders]. """
Output: """ The human mind can be altered by any number of substances. These may be medications used to treat illness, infection, or even some of the mental disorders that we've looked at elsewhere in this series. They can be components of the food or beverages we consume that may change our cognition or behaviour, or they can even be substances used to intentionally affect our behaviour or perceptions of the world. However, any number of mind-altering substances can be used to the point that they cause clinically significant distress or functional impairment, leading to possible intoxication, abuse, addiction and withdrawal. This is the case for those who experience Addictive Disorders, also known as Substance-related and Addictive Disorders. These individuals may have their cognition, perceptions or physiology debilitated in reversible or irreversible ways. They may continue to use these substances, even when they experience clusters of negative effects. And they may become so dependent on a substance that even when they try to cease or reduce their use, they experience additional behavioral or cognitive consequences. Substance-related disorders are part of a larger grouping of disorders called substance-related and addictive disorders codified and defined in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders or DSM-5. These disorders share a common trait of activating the brain's reward system by a substance or addiction, and the individuals with these disorders continue their use and behavior, whether or not they recognize it as a problem. """	**Output:** """ The human mind can be influenced by various substances, including medications used for medical treatment or components of food and beverages, leading to alterations in cognition and behavior. Additionally, substances may be intentionally used to affect behavior or perceptions. However, excessive use of mind-altering substances can result in clinically significant distress and functional impairment, potentially leading to intoxication, abuse, addiction, and withdrawal. Addictive Disorders, also known as Substance-related and Addictive Disorders, encompass conditions where individuals experience debilitation of cognition, perceptions, or physiology, either in reversible or irreversible ways. Despite experiencing negative effects, individuals with Addictive Disorders may persist in substance use and face additional behavioral or cognitive consequences when attempting to reduce or cease usage. These disorders are part of a broader classification known as substance-related and addictive disorders, defined in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). A common characteristic of these disorders is the activation of the brain's reward system through substance use or addiction, irrespective of the individual's recognition of the issue. To effectively address and alleviate Addictive Disorders, consider the following professional suggestions: 1. Seek specialized help: Encourage individuals struggling with Addictive Disorders to seek guidance from healthcare professionals with expertise in addiction treatment. Qualified professionals can conduct comprehensive assessments, accurately diagnose the condition, and design personalized treatment plans. 2. Adopt a comprehensive approach: Employ a comprehensive treatment approach that integrates evidence-based therapeutic interventions and, if appropriate, medication. Tailor the treatment plan to address the specific substance(s) involved and meet the unique needs of each individual. 3. Individualized therapy: Engage individuals in individual therapy, such as cognitive-behavioral therapy (CBT) or interpersonal therapy (IPT), to explore and address the underlying factors contributing to addictive behaviors, develop effective coping strategies, and promote positive behavioral changes. 4. Medication-assisted treatment (MAT): Consider medication-assisted treatment in conjunction with therapy for certain Addictive Disorders. MAT can aid in managing cravings, reducing withdrawal symptoms, and supporting sustained recovery. 5. Support groups and peer assistance: Encourage participation in support groups or group therapy to foster a sense of community, empathy, and shared experiences among individuals facing similar challenges. 6. Dual diagnosis treatment: For individuals with co-occurring mental health conditions, prioritize dual diagnosis treatment that simultaneously addresses the addiction and underlying mental health issues, promoting holistic recovery. 7. Involve family support: Engage family members in the treatment process to provide understanding and support. Family therapy can enhance communication and create a supportive environment conducive to recovery. 8. Develop relapse prevention strategies: Collaborate with healthcare professionals to develop personalized relapse prevention strategies, identifying triggers and coping mechanisms to minimize the risk of relapse. 9. Emphasize lifestyle changes: Promote a healthy lifestyle that includes regular exercise, proper nutrition, sufficient sleep, stress management techniques, and engaging in activities that support overall well-being. 10. Sustain aftercare support: Recognize that recovery from Addictive Disorders is an ongoing process. Offer continued support, including regular follow-up appointments, therapy sessions, and participation in support groups to sustain recovery and address any challenges that may arise. Remember, the journey to recovery from Addictive Disorders is unique for each individual. Collaborating closely with qualified healthcare professionals ensures a tailored approach that addresses their specific needs, fostering lasting recovery and improved well-being. """
Task: """ Question answering task. """	**Task:** """ Question answering task. """
Domain: """ Addictive Disorders. """	**Domain:** """ Addictive Disorders. """

Table 6.10: The Zero-Shot, Natural-Instruction and Assistant-Instruction Tuned ChatGLM-6B on Psychotherapy data.

The Zero-Shot Tuned ChatGLM-6B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Bipolar Disorders]. """
Input:
""" We are discussing [Bipolar Disorders]. """
Output:
"""
Thank you for bringing up Bipolar Disorder. Here are some suggestions that may be helpful:
1. Seek professional help: If you are experiencing symptoms of Bipolar Disorder, it is important to seek help from a mental health professional. They will be able to provide you with the tools and support you need to manage your symptoms and improve your overall well-being. 2. Practice self-care: Bipolar Disorder can be a challenging condition to manage, but with the right support and self-care, it is possible to alleviate symptoms and improve quality of life. This may include activities such as exercise, healthy eating, and getting enough rest. 3. Follow a consistent routine: Consistency is key when it comes to managing Bipolar Disorder. Try to stick to a regular routine as much as possible, including your sleep schedule, medication regimen, and any other triggers that may contribute to your symptoms. 4. Learn to manage stress: Stress can exacerbate symptoms of Bipolar Disorder, so it is important to learn ways to manage stress and prevent it from triggering your symptoms. This may include therapy, meditation, or other relaxation techniques. 5. Consider medication: In some cases, medication may be necessary to help manage symptoms of Bipolar Disorder. It is important to work closely with a mental health professional to determine the right dosage and treatment plan for you.
I hope these suggestions are helpful. Let me know if you have any other questions or concerns.
"""

The constructed Natural-Instruction Tuned ChatGLM-6B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Bipolar Disorders]. """
Input:
""" We are discussing [Bipolar Disorders]. """
Output:
"""
Thank you for bringing up Bipolar Disorder. Here are some suggestions that may be helpful:
1. Seek professional help: It's important to consult with a mental health professional who specializes in Bipolar Disorder. They will be able to provide you with the proper treatment and medication. 2. Practice self-care: Engage in activities that bring you joy and help you relax. This can include things like exercise, meditation, or hobbies. 3. Follow a consistent routine: Maintain a regular sleep schedule, eat a healthy diet, and try to stick to a routine as much as possible. 4. Connect with others: Reach out to friends and family members for support. 5. Avoid triggers: Identify things that trigger your Bipolar symptoms and try to avoid them.
Remember, it's important to seek professional help and work closely with a mental health professional to manage your symptoms and improve your overall well-being.
"""

The constructed Assistant-Instruction Tuned ChatGLM-6B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Bipolar Disorders]. """
Input:
""" We are discussing [Bipolar Disorders]. """
Output:
"""
Thank you for bringing up Bipolar Disorder. Here are some suggestions that may be helpful:
1. For individuals with Bipolar Disorder, it is important to seek professional help as soon as possible. This can include seeking assistance from a mental health professional, such as a psychiatrist or psychologist, who can provide appropriate treatment and support. 2. It is also important for individuals with Bipolar Disorder to establish a regular routine and stick to it. This can include setting aside time for exercise, maintaining a consistent sleep schedule, and following a daily routine that is consistent with medication. 3. In addition, it can be helpful for individuals with Bipolar Disorder to engage in activities that promote relaxation and reduce stress. This can include hobbies or creative pursuits, such as painting or playing music, or spending time with friends or loved ones. 4. It is also important for individuals with Bipolar Disorder to inform their loved ones and family members about their condition. This can help to reduce stigma and provide support and understanding. 5. Finally, it is important for individuals with Bipolar Disorder to seek professional help when they are experiencing a manic episode or a depressive episode. This can include seeking assistance from a mental health professional or calling emergency services if necessary.
I hope these suggestions are helpful. Let me know if you have any further questions or concerns.
"""

Table 6.11: The Zero-Shot, Natural-Instruction and Assistant-Instruction Tuned MPT-7B on Psychotherapy data.

The Zero-Shot Tuned MPT-7B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Addictive Disorders]. """
Input:
""" We are discussing [Addictive Disorders]. """
Output:
"""
Addictive disorders are a group of mental health conditions characterized by compulsive drug seeking and use despite harmful consequences. The most common addictive disorders are alcoholism, drug addiction, and gambling addiction.
Addictive disorders are often chronic, relapsing, and progressive. They can be caused by a combination of genetic, environmental, and psychological factors.
Treatment for addictive disorders typically involves a combination of behavioral therapy, medication, and support groups.
Addictive disorders can be difficult to treat, but there are many resources available to help people overcome their addiction.
If you or someone you know is struggling with an addictive disorder, please seek help. There is no shame in asking for help.
"""

The constructed Natural-Instruction Tuned MPT-7B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Addictive Disorders]. """
Input:
""" We are discussing [Addictive Disorders]. """
Output:
"""
Addictive disorders are a group of mental health conditions characterized by compulsive drug seeking and use despite harmful consequences. The most common addictive disorders are alcoholism, drug addiction, gambling addiction, and internet addiction.
Addictive disorders are classified as substance-related disorders in the Diagnostic and Statistical Manual of Mental Disorders (DSM). They are also classified as addictive, impulse-control, and related disorders in the International Classification of Diseases (ICD).
Addictive disorders are often chronic and progressive. They can cause significant impairment in social, occupational, and other important areas of functioning.
Addictive disorders are treatable. Treatment may include medications, behavioral therapy, and/or support groups.
"""

The constructed Assistant-Instruction Tuned MPT-7B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Addictive Disorders]. """
Input:
""" We are discussing [Addictive Disorders]. """
Output:
"""
I think it's important to understand what addiction is, and how it differs from other types of compulsive behavior. Addiction is a chronic brain disease that causes compulsive drug seeking and use, despite harmful consequences. It affects the brain's reward system, which is responsible for motivating us to seek out pleasurable activities. When someone becomes addicted, their brain is wired to seek out drugs, even when they know they're harmful. This can lead to risky behaviors, such as driving while intoxicated, or engaging in sexual activity with multiple partners.
Addiction is a complex disease that requires a multifaceted approach to treatment. Treatment should include both medical and psychological components, as well as support from family and friends. Medical treatment may include medications to help reduce cravings and ease withdrawal symptoms. Psychological treatment may include cognitive behavioral therapy (CBT), which helps people change their thinking patterns and behaviors. Support from family and friends is essential for recovery, as it provides a sense of stability and accountability.
"""

Table 6.12: The Zero-Shot, Natural-Instruction and Assistant-Instruction Tuned Falcon-7B on Psychotherapy data.

The Zero-Shot Tuned Falcon-7B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Depressive Disorders]. """
Input:
""" We are discussing [Depressive Disorders]. """
Output:
"""
Depressive Disorders are a serious mental health condition that can significantly impact an individual's daily life. It is important to seek professional help if you or someone you know is experiencing symptoms such as persistent sadness, loss of interest in activities, changes in appetite or sleep patterns, and difficulty concentrating. A mental health professional can provide effective treatment options, including therapy and medication, to help alleviate symptoms and improve overall well-being. It is also important to maintain a healthy lifestyle, including regular exercise, a balanced diet, and social connections, to support mental health.
"""

The constructed Natural-Instruction Tuned Falcon-7B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Depressive Disorders]. """
Input:
""" We are discussing [Depressive Disorders]. """
Output:
"""
Depressive Disorders are a serious mental health condition that can significantly impact an individual's daily life. It is important to seek professional help if you or someone you know is experiencing symptoms such as persistent sadness, loss of interest in activities, changes in appetite or sleep patterns, and difficulty concentrating. A mental health professional can provide effective treatment options, including therapy and medication, to alleviate symptoms and improve overall well-being. It is also important to create a support network of friends and family to provide emotional support during this time.
"""

The constructed Assistant-Instruction Tuned Falcon-7B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Depressive Disorders]. """
Input:
""" We are discussing [Depressive Disorders]. """
Output:
"""
Depressive Disorders are a complex mental health condition that requires a multifaceted approach to effectively address and alleviate. Professional suggestions may include seeking professional help from a mental health professional, engaging in regular physical activity, and cultivating a strong support network. It is also important to maintain a healthy diet and practice self-care. Engaging in activities that bring joy and fulfillment can also be helpful in managing symptoms. Therapy is also a highly effective treatment for Depressive Disorders. Therapy can help individuals learn new coping strategies and develop a more positive outlook. It is important to seek professional help from a licensed therapist to determine the best course of treatment for each individual.
"""

Table 6.13: The Zero-Shot, Natural-Instruction and Assistant-Instruction Tuned Llama2-7B on Psychotherapy data.

The Zero-Shot Tuned Llama2-7B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Autism]. """
Input:
""" We are discussing [Autism]. """
Output:
"""
Autism is a developmental disorder that can cause significant social, communication and behavioral challenges. There is no known single cause for autism, but increased awareness and funding support research into possible causes and better ways to treat the disorder.
"""

The constructed Natural-Instruction Tuned Llama2-7B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Autism]. """
Input:
""" We are discussing [Autism]. """
Output:
"""
Autism is a developmental disorder that can cause significant social, communication and behavioral challenges. It is a spectrum disorder, which means that symptoms and severity vary widely among individuals.
"""

The constructed Assistant-Instruction Tuned Llama2-7B on Psychotherapy data.

Instruction:
""" Please provide professional suggestions or comments to effectively address and alleviate [Autism]. """
Input:
""" We are discussing [Autism]. """
Output:
"""
Autism is a developmental disorder that can cause significant social, communication and behavioral challenges. It is a spectrum disorder, which means that symptoms can range from mild to severe. Some people with autism may also have other conditions, such as attention deficit hyperactivity disorder (ADHD), anxiety disorders, or epilepsy. Applied Behavior Analysis (ABA) a widely recognized and evidence-based therapy for Autism. ABA techniques focus on promoting positive behaviors and reducing challenging behaviors. Speech and language therapy can help improve communication skills, such as speech articulation, expressive language, and social communication.
"""

Table 6.14: Two examples of using Falcon-7B. $M1$, $M2$ and $M3$ respectively is the Readability, Professional and Match Score. For example, 4/5 means Readability of original LLM / Readability of finetuned-LLM.

Original LLM	Assistant Instruction Fine-Tuned LLM	M1	M2	M3	Comments
Question: What is Depressive Disorders, and what is its symptoms?					
Depressive disorders are a group of mental health conditions characterized by persistent feelings of sadness, loss of interest in activities, and changes in mood and behavior. Symptoms may include feelings of worthlessness, guilt, and hopelessness, as well as changes in appetite, sleep patterns, and energy levels. It is important to seek professional help if you or someone you know is experiencing these symptoms.	Depressive disorders are a group of mental health conditions characterized by persistent feelings of sadness, loss of interest in activities, and changes in mood and behavior. Symptoms may include feelings of worthlessness, guilt, and hopelessness, as well as changes in appetite, sleep patterns, and energy levels. It is important to seek professional help if you or someone you know is experiencing these symptoms.	5/5	5/5	1/1	This is really good.
Question: What is Depression Behavioral Activation and Cognitive Change?					
Depression Behavioral Activation and Cognitive Change (DBAC) is a cognitive behavioral therapy (CBT) approach that aims to help individuals with depression by addressing both the behavioral and cognitive aspects of the condition. It involves identifying and changing negative thought patterns and behaviors, developing coping strategies, and improving overall mood. DBAC has been shown to be effective in reducing symptoms of depression and improving quality of life for those who undergo it.	Depression Behavioral Activation and Cognitive Change (DBAC) is a cognitive behavioral therapy (CBT) approach that aims to help individuals with depression by addressing both the behavioral and cognitive aspects of the condition. It involves identifying and changing negative thought patterns and behaviors, as well as developing new coping strategies and problem-solving skills. DBAC has been shown to be effective in reducing symptoms of depression and improving overall quality of life.	3/5	4/5	1/1	.

Table 6.15: One example of using Llama2-7B. $M1$, $M2$ and $M3$ respectively is the Readability, Professional and Match Score (original LLM / finetuned-LLM).

Original LLM	Assistant Instruction Fine-Tuned LLM	M1	M2	M3	Comments
Question: What is Depressive Disorders, and what is its symptoms?					
Depressive disorder is a mental illness that causes a persistent feeling of sadness and loss of interest. It can interfere with daily life, work, school, and relationships. The symptoms of depressive disorder include: - Persistent sad, anxious, or empty mood - Feelings of hopelessness, pessimism, guilt, worthlessness, helplessness - Loss of interest or pleasure in activities that were once enjoyed, including sex - Decreased energy, fatigue, being ""slowed down"" - Difficulty concentrating, remembering details, making decisions - Insomnia, early-morning wakefulness, or excessive sleeping - Overeating, or appetite loss - Thoughts of death or suicide, or suicide attempts ###Reference: https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007.	Depressive disorder is a mental illness that causes a persistent feeling of sadness and loss of interest. It can interfere with daily life, causing severe emotional and physical problems. Depression is a common but serious mood disorder. It causes severe symptoms that affect how you feel, think, and handle daily activities, such as sleeping, eating, or working. To be diagnosed with depression, the symptoms must be present for at least two weeks.	4/4	3/3	1/2	More examples of symptoms would be better. It is tricky to say "To be diagnosed with depression, the symptoms must be present for at least two weeks." because it is not the only one criterion for diagnosis.

Table 6.16: One example of using ChatGLM2-6B. $M1$, $M2$ and $M3$ respectively is the Readability, Professional and Match Score (original LLM / finetuned-LLM).

Original LLM	Assistant Instruction Fine-Tuned LLM	M1	M2	M3	Comments
Question: What is Depressive Disorders, and what is its symptoms?					
Some suggestions or comments that could address or alleviate the instruction's input could include: 1. Depressive disorders are a serious mental health condition that can affect many aspects of a person's life. Symptoms may include persistent feelings of sadness or hopelessness, loss of interest in activities, changes in appetite or sleep patterns, and low energy. It is important to seek professional help if these symptoms are affecting daily life. 2. Addressing depressive disorders requires a combination of medication and therapy. Medications such as selective serotonin reuptake inhibitors (SSRIs) and non-steroidal anti-inflammatory drugs (NSAIDs) can help alleviate symptoms while therapy can help identify and address the underlying causes of the depression. It is important to work with a mental health professional to find the right treatment plan. 3. It is important to note that everyone's experience with depressive disorders is different, and seeking professional help is a journey and not a one-time event. It is also important to be patient and understanding with oneself throughout the process. 4. In addition to professional help, there are many things that can be done to alleviate symptoms of depressive disorders, such as regular exercise, healthy eating, and getting enough sleep. It is important to prioritize self-care and seek support from loved ones or a therapist if needed. 5. It is important to remember that depressive disorders are a treatable condition, and with the right treatment and support, recovery is possible. Overall, addressing depressive disorders requires a multi-faceted approach that includes medication, therapy, and self-care. It is important to seek professional help and work with a mental health professional to find the right treatment plan.	Certainly, I'd be happy to provide some information and answer any questions you may have about depressive disorders and their symptoms. Depressive disorders are a common mental health condition that refers to a range of symptoms that affect a person's ability to function and maintain a healthy lifestyle. Some of the most common symptoms of depressive disorders include: 1. Persistent feelings of sadness or hopelessness: This involves a persistent feeling of sadness or hopelessness that can be associated with loss, rejection, or other negative events in life. 2. Loss of interest in activities that were previously enjoyed: This involves a loss of interest in activities that were previously enjoyable or fulfilling. 3. Changes in sleep patterns: This involves changes in the patterns of sleep that can result in insomnia, difficulty sleeping, or sleeping too much. 4. Changes in appetite: This involves changes in appetite that can result in weight gain or loss. 5. Difficulty concentrating: This involves difficulty concentrating on tasks or ideas, which can have a significant impact on a person's ability to function in their daily life. 6. Thoughts of self-harm or suicide: This involves thoughts of self-harm or suicide, which can be a symptom of depression. It's important to note that these symptoms can vary from person to person, and some people may experience different symptoms or symptoms to those listed above. Additionally, not everyone who experiences these symptoms will necessarily have a depressive disorder. However, if you are experiencing any of these symptoms and they are causing significant distress or impacting your daily life, it's important to seek professional help. It's also important to seek help early on, as untreated depression can have serious consequences for both physical and mental health. Some common treatments for depressive disorders include therapy, medication, and lifestyle changes, such as regular exercise, healthy eating, and getting enough sleep.	4/5	3/5	2/1	It is more on point, without confabulations

Table 6.17: Two examples of using MPT-7B. $M1$, $M2$ and $M3$ respectively is the Readability, Professional and Match Score (original LLM / finetuned-LLM).

Original LLM	Assistant Instruction Fine-Tuned LLM	M1	M2	M3	Comments
Question: What is Depressive Disorders, and what is its symptoms?					
Depressive disorders are a group of mood disorders characterized by depressed mood, loss of interest or pleasure in activities, weight loss or gain, insomnia or hypersomnia, fatigue, feelings of worthlessness or guilt, slowed thinking, and recurrent thoughts of death or suicide. Depressive disorders are classified as major depressive disorder (MDD), dysthymia, bipolar disorder, and postpartum depression. MDD is the most common type of depressive disorder. It is characterized by depressed mood, loss of interest or pleasure in activities, weight loss or gain, insomnia or hypersomnia, fatigue, feelings of worthlessness or guilt, slowed thinking, and recurrent thoughts of death or suicide. Dysthymia is a chronic form of depression that is less severe than MDD but more severe than normal sadness. Bipolar disorder is a mental disorder that causes extreme shifts in mood, energy, activity levels, and the ability to carry out day-to-day tasks. Postpartum depression is a type of depression that occurs after childbirth. Symptoms of depressive disorders include: - Depressed mood - Loss of interest or pleasure in activities - Weight loss or gain - Insomnia or hypersomnia - Fatigue - Feelings of worthlessness or guilt - Slowed thinking - Recurrent thoughts of death or suicide	Depressive disorders are a group of mood disorders characterized by depressed mood, loss of interest in activities, feelings of worthlessness or guilt, disturbed sleep or appetite, fatigue, and sometimes suicidal thoughts.	3/4	3/3	2/3	The comment of fine-tuned LLM is too short and brief.
Question: What is Depression Behavioral Activation and Cognitive Change?					
Depression Behavioral Activation and Cognitive Change is a treatment for depression that focuses on helping people with depression to change their behavior and thinking patterns. The goal of this treatment is to help people with depression to become more active, to think more positively about themselves and their lives, and to make changes in their lives that will help them to feel better. The treatment involves working with a therapist to identify areas where people with depression can make changes in their lives. The therapist will help people with depression to set goals for making these changes and to develop a plan for how they will make these changes. The therapist will also help people with depression to identify and overcome any obstacles that might get in the way of making these changes.	Depression Behavioral Activation and Cognitive Change is a treatment for depression that focuses on helping people change their behavior and thinking patterns. It is based on the idea that depression is caused by negative thoughts and behaviors, and that changing these thoughts and behaviors can help people feel better. The treatment involves working with a therapist to identify negative thoughts and behaviors and then coming up with ways to change them. The therapist will help the person come up with new ways of thinking and behaving that are more positive and helpful. The treatment is based on the idea that depression is caused by negative thoughts and behaviors, and that changing these thoughts and behaviors can help people feel better.	5/3	5/5	1/1	There are some repetitive information.

Chapter 7

Conclusion

7.1 Summary

Depression is a complex, multi-symptomatic and highly recrudescent mental disease. Severity detection and psychotherapy have only stated to be explored. In this work, we answered the big questions about scoring depressive severity and how to provide universal psychotherapy to depressive patients. In chapter 2, we found increased delta deactivation accompanied by strong beta activation is the main feature of depression as the depression becomes more severe. We also verified that the ANN models using EEGs can detect depression and depressive severity. In Chapter 3, we found that detecting depressive severity is not a pure classification or scoring task. Therefore, we designed a fuzzy window with Gaussian processed labels method that can tackle the ordinal scoring tasks, such as depressive severity. In Chapter 4, we found that the bilateral PFC mainly focused on some cognitive components, for example, rehearsal before recognition to classify objects, inhibition to maintain positive memory and activities, and disinhibition to arouse or activate subsequent interactions in brain. Meanwhile, the right PFC sometimes could assist left PFC to implement high capacity WM tasks. By contrast, the posterior regions, PPC, tends to be engaged in attention arousing and maintaining. These two findings suggest that **a)** the recurrent maintenance circuit may keep the brain executing positive cognitive components, **b)** then the instantly monitoring inhibition would pause the deadlocked sustention function to save energy, and **c)** the arriving of disinhibition arouses the next step in brain to select new subject or focus on novel subjects. In Chapter 5, we answered questions on how pretrained language models can improve their performance on fine-tuning downstream tasks by purely applying the mutation of the attention block in Transformers, and found further evidence that inhibited gate MLPs mechanism is important to fine-tune language downstream tasks. In Chapter 6, we observed that pre-trained LLMs fine-tuned on Psychotherapy Assistant Instructions outperform SOTA LLMs response baselines. Our Assistant-Instruction approach offers a half-annotation method to align pre-trained LLMs with instructions. We also released our large synthetic dataset, facilitating future studies on professional instruction tuning

7.2 Contributions and Achievements

Scientific contributions of this Thesis is represented by the following achievements:

1. In Chapter 2, we presented the central-parietal increased delta deactivation accompanied by strong beta activation in the severe depression group under working memory tasks. We also proposed models with specific frequencies and brain regions for detecting depression and scoring the depressive severity based on two professional psychologists' score labels. These findings were published on IEEE Transactions on Neural Systems and Rehabilitation Engineering (IF: 4.9).

2. In Chapter 3, we proposed one method to reduce the influence of the overlapping features among the ordinal neighbor classes. This process can effectively improve the scoring performance of the ordinal classification. When the ordinal sequence of the labels or images is not consecutive, FW-GPL can achieve an equivalent performance to wholly sequential ordinal data by setting a proper length for the fuzzy window. We published this method on Applied Sciences (IF: 2.7).

3. In Chapter 4, we examined brain networks of WM by phase-lock coherence and directional coherence after the 64 channels EEG adaptively fitted and four sources generated to simulate cerebral internal communications. We proposed our "neurocognitive architecture" of WM based on region-to-region connections, and found the pathways of memory maintenance and lateral inhibition during WM. We published these findings on IEEE Transactions on Neural Systems and Rehabilitation Engineering (IF: 4.9, WOS citations: 14).

4. In Chapter 5, we proposed one adaption fine-tuning methond - InA that can effectively inhibit irrelevant information during fine-tuning on downstream tasks, enabling the model to focus more on task-related information and eliminating the impact of irrelevant knowledge. InA gains the capability to suppress task-irrelevant knowledge by subtracting a threshold. We have submitted this method to Neural Networks (Minor Revision).

5. In Chapter 6, we released psychotherapy data, revised by GPT-4. This process enhanced the LLMs' understanding of new professional knowledge and enables them to generate content close to GPT-4.This chapter demonstrated the effectiveness of

using GPT-4-revised data for LLM instruction-tuning, providing practical insights for building a general-purpose instruction-following agent powered by LLMs (e.g., GPT-4). We have submitted this method to Knowledge-Bsed System (Major Revision).

7.3 Future Work

In my future work, there are five main directions as follow.

- to make the close loop of Figure 1.1 more stable, practical and convenient. For example, to enlarge the clinical data pool, to optimise the models based on the feedback of experts, as well as to improve the adaption and deployability of large language models on psychotherapy domains.
- to develop advanced parameter-efficient fine-tuning methods that have a faster inference speed and smaller tunable parameters.
- to improve the performance of psychotherapy aiding chatbot on more domains. For example, auxiliary diagnosis, support of treatment comments, and diary emotion monitoring.
- to develop bipolar disorders detection system using brain computer interface systems.
- to develop control algorithms that can control the learning systems of most ANN models, especially on the application of brain computer interface and large language models.

List of Candidate's Publications Related to the Thesis

7.4 Publications in Impacted Journals

This thesis builds on the results previously published in the following publications:

1. **Kang, C.***; Li, Y.*; Novak, D.; Zhang, Y.; Zhou, Q.; Hu, Y. (2020). Brain Networks of Maintenance, Inhibition and Disinhibition During Working Memory. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 7, pp. 1518-1527, July 2020.

2. **Kang, C.**; Novak, D.; Yao, X.; Xie, J.; Hu, Y. (2023). Classifying and Scoring Major Depressive Disorders by Residual Neural Networks on Specific Frequencies and Brain Regions. IEEE Trans Neural Syst Rehabil Eng. 2023;31:2964-2973.

3. **Kang, C.**; Yao, X.; Novak, D. (2023). Fuzzy Windows with Gaussian Processed Labels for Ordinal Image Scoring Tasks. Appl. Sci. 2023, 13, 4019.

4. **Kang, C.**; Prokop, J.; Tong, L.; Zhou, Z.; Hu, Y.; Novak, D. (2023). InA: Inhibition Adaption On Pre-trained Language Models. Submitted to Neural Networks (Minor Revision)

5. **Kang, C.**; Cheng, Y.; Zhang, Y.; Hu, Y.; Novak, D. (2023). Domain Specific Assistant Instruction on Psychotherapy Chatbot. Submitted to Knowledge-Based Systems (Major Revision)

The following publications are related to the topic but were not included in the thesis, in order to keep the thesis more focused and easier to follow:

1. **Kang, C.**, Yu, X., Wang, S. H., Guttery, D., Pandey, H., Tian, Y., Zhang, Y. (2020). A heuristic neural network structure relying on fuzzy logic for images scoring. IEEE Transactions on Fuzzy Systems., vol. 29, no. 1, pp. 34-45, Jan. 2021, doi: 10.1109/TFUZZ.2020.2966163.

2. Li, Y.*, **Kang, C.***, Wei, Z., Qu, X., Liu, T., Zhou, Y., Hu, Y. (2017). Beta oscillations in major depression – signaling a new cortical circuit for central executive function. Scientific reports, 7 (1), 1-15, doi: 10.1038/s41598-017-18306-w.

3. Li, Y.*, **Kang, C.***, Qu, X., Zhou, Y., Wang, W., Hu, Y. (2016). Depression-related brain connectivity analyzed by EEG event-related phase synchrony measure. Frontiers in human neuroscience, 10, 477, doi: 10.3389/fnhum.2016.00477.

4. Cui, H., Li, H., Li, G., **Kang, C.**, Yao, X., Feng, S., Hu, Y. (2019). Utility of trial-to-trial latency variability of somatosensory evoked potentials for diagnosis of spinal cord demyelination. Journal of neurotrauma, 36(24), 3356-3362, doi: 10.1089/neu.2018.6293.

7.5 Other Publications

The following publications were published during the duration of the Ph.D. but are not included in the thesis because they are not directly related to the topic of the thesis:

1. Yu, X., **Kang, C.**, Guttery, DS, Kadry, S., Chen, Y., Zhang, Y. (2020). ResNet-SCDA-50 for breast abnormality classification. IEEE / ACM transactions on computational biology and bioinformatics, vol. 18, no. 1, pp. 94-102, 1 Jan.-Feb. 2021, doi: 10.1109/TCBB.2020.2986544.

2. Yao, X., Zhu, Z., **Kang, C.**, Wang, S., Gorriz, J., Zhang, Y. (2022). AdaD-FNN for Chest CT-Based COVID-19 Diagnosis. IEEE Transactions on Emerging Topics in Computational Intelligence, doi: 10.1109/TETCI.2022.3174868.

The following publications were not included as they are currently under review:

1. Guo, Z.*, Chen, Y. *, **Kang, C.** (2024). Unipolar and Bipolar Disorders Classification: Facial-Based EEGs and DCNN.

2. Wang, H., Wen, J., **Kang, C.** (2024). Abnormal Brain EEG Networks in Bipolar Disorders under Facial-Emotional Experiments.

3. **Kang, C.**, Prokop, J., Tong, L., Zhou, H., Novak, D. (2024). Gate and Inhibition Mechanism in MLPs and BERTs.

4. Yao, X.*, **Kang, C.***, Zhang, X., Wang, S., Zhang, Y. (2024). FuzH-PID: Highly Controllable and Stable DNN for COVID-19 Detection via Improved Stochastic Optimization.

5. **Kang, C.***, Yao, X.* (2023). Based on What We Can Control Artificial Neural Networks.

6. **Kang, C.**, Chen, X., Hu, Y., Novak, D. (2024). Quantized Embedding Vectors for Controllable Diffusion Language Models.

7. **Kang, C.**, Yao, X., Prokop, J., Tong, L., Zhou, H., Hu, Y., Novak, D. (2024). Shunting Inhibition on Artificial Neural Networks.

Bibliography

[1] C. Kang, Y. Li, D. Novak, Y. Zhang, Q. Zhou, and Y. Hu, "Brain networks of maintenance, inhibition and disinhibition during working memory", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 7, pp. 1518–1527, 2020. DOI: 10.1109/TNSRE.2020.2997827.

[2] C. Kang, X. Yu, S.-H. Wang, *et al.*, "A heuristic neural network structure relying on fuzzy logic for images scoring", *IEEE transactions on fuzzy systems*, vol. 29, no. 1, pp. 34–45, 2020.

[3] X. Yu, C. Kang, D. S. Guttery, S. Kadry, Y. Chen, and Y.-D. Zhang, "Resnet-scda-50 for breast abnormality classification", *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 1, pp. 94–102, 2020.

[4] X. Yao, Z. Zhu, C. Kang, S.-H. Wang, J. M. Gorriz, and Y.-D. Zhang, "Adad-fnn for chest ct-based covid-19 diagnosis", *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 1, pp. 5–14, 2022.

[5] C. Kang, X. Yao, and D. Novak, "Fuzzy windows with gaussian processed labels for ordinal image scoring tasks", *Applied Sciences*, vol. 13, no. 6, p. 4019, 2023.

[6] C. Kang, D. Novák, X. Yao, J. Xie, and Y. Hu, "Classifying and scoring major depressive disorders by residual neural networks on specific frequencies and brain regions", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 2964–2973, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259369369.

[7] Y. Li, C. Kang, X. Qu, Y. Zhou, W. Wang, and Y. Hu, "Depression-related brain connectivity analyzed by eeg event-related phase synchrony measure", *Frontiers in human neuroscience*, vol. 10, p. 477, 2016. DOI: https://doi.org/10.3389/fnhum.2016.00477.

[8] Y. Li, C. Kang, Z. Wei, *et al.*, "Beta oscillations in major depression–signalling a new cortical circuit for central executive function", *Scientific reports*, vol. 7, no. 1, pp. 1–15, 2017. DOI: https://doi.org/10.1038/s41598-017-18306-w.

[9] C. Kang, J. Prokop, L. Tong, H. Zhou, Y. Hu, and D. Novak, "Gimlps: Gate with inhibition mechanism in mlps", *arXiv preprint arXiv:2208.00929*, 2022.

[10] G. Darroch. "Netherlands 'will pay the price' for blocking turkish visit – erdoğan". (2017), [Online]. Available: `https://www.theguardian.com/world/2017/mar/12/netherlands-will-pay-the-price-for-blocking-turkish-visit-erdogan` (visited on 03/12/2017).

[11] S. Evans-Lacko, S. Aguilar-Gaxiola, A. Al-Hamzawi, *et al.*, "Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the who world mental health (wmh) surveys", *Psychological medicine*, vol. 48, no. 9, pp. 1560–1571, 2018.

[12] G. V. P. Reddy, "Depression – the global crisis", vol. 34, no. 3, p. 201, 2013.

[13] B. F. Grant, F. S. Stinson, D. A. Dawson, *et al.*, "Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: Results from the national epidemiologic survey on alcohol and related conditions", vol. 29, no. 7, pp. 807–16, 2006.

[14] M. Vermani, M. Marcus, and M. A. Katzman, "Rates of detection of mood and anxiety disorders in primary care: A descriptive, cross-sectional study", vol. 13, no. 2, 2011.

[15] J. C. Fournier, N. R. Forand, Z. Wang, *et al.*, "Initial severity and depressive relapse in cognitive behavioral therapy and antidepressant medications: An individual patient data meta-analysis", *Cognitive Therapy and Research*, vol. 46, no. 3, pp. 517–531, 2022, ISSN: 1573-2819.

[16] I. Kirsch, B. J. Deacon, T. B. Huedo-Medina, A. Scoboria, T. J. Moore, and B. T. Johnson, "Initial severity and antidepressant benefits: A meta-analysis of data submitted to the food and drug administration", *PLoS medicine*, vol. 5, no. 2, e45, 2008, ISSN: 1549-1277.

[17] H. S. Sharma, M. Chopp, L. Chen, *et al.*, "The 2021 yearbook of neurorestoratology", *Journal of Neurorestoratology*, p. 100 008, 2022.

[18] J. Davies and J. Read, "A systematic review into the incidence, severity and duration of antidepressant withdrawal effects: Are guidelines evidence-based?", *Addictive behaviors*, vol. 97, pp. 111–121, 2019, ISSN: 0306-4603.

[19] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated eeg-based screening of depression using deep convolutional neural network", *Computer methods and programs in biomedicine*, vol. 161, pp. 103–113, 2018, ISSN: 0169-2607.

[20] G. Andrews and L. Peters, "The psychometric properties of the composite international diagnostic interview", vol. 33, no. 2, pp. 80–88,

[21] M. Buoli, B. M. Cesana, J. L. Barkin, G. Tacchini, and A. C. Altamura, "Validity of a clinical diagnosis of bipolar disorder among participants in a multicenter study using the mini-international neuropsychiatric interview", *Bipolar disorders*, vol. 20, no. 3, pp. 284–284, 2018, ISSN: 1398-5647.

[22] B. K. Natamba, J. Achan, A. Arbach, *et al.*, "Reliability and validity of the center for epidemiologic studies-depression scale in screening for depression among hiv-infected and-uninfected pregnant women attending antenatal services in northern uganda: A cross-sectional study", *BMC psychiatry*, vol. 14, no. 1, pp. 1–8, 2014, ISSN: 1471-244X.

[23] M. B. First and M. Gibbon, "The structured clinical interview for dsm-iv axis i disorders (scid-i) and the structured clinical interview for dsm-iv axis ii disorders (scid-ii)", 2004. [Online]. Available: `https://psycnet.apa.org/record/2004-12821-011`.

[24] M. Zimmerman, J. H. Martinez, D. Young, I. Chelminski, and K. Dalrymple, "Severity classification on the hamilton depression rating scale", *Journal of Affective Disorders*, vol. 150, no. 2, pp. 384–388, 2013, ISSN: 0165-0327. DOI: `https://doi.org/10.1016/j.jad.2013.04.028`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0165032713003017`.

[25] G. Jackson-Koku, "Beck depression inventory", *Occupational Medicine*, vol. 66, no. 2, pp. 174–175, 2016.

[26] M. Mousavian, J. Chen, Z. Traylor, and S. Greening, "Depression detection from smri and rs-fmri images using machine learning", *Journal of Intelligent Information Systems*, vol. 57, no. 2, pp. 395–418, 2021.

[27] R. Wang, Y. Hao, Q. Yu, M. Chen, I. Humar, and G. Fortino, "Depression analysis and recognition based on functional near-infrared spectroscopy", *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 12, pp. 4289–4299, 2021.

[28] Y. Zhu, J. K. Jayagopal, R. K. Mehta, *et al.*, "Classifying major depressive disorder using fnirs during motor rehabilitation", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 4, pp. 961–969, 2020.

[29] T. Roh, S. Hong, and H.-J. Yoo, "Wearable depression monitoring system with heart-rate variability", in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2014, pp. 562–565.

[30] J. Shen, X. Zhang, X. Huang, *et al.*, "An optimal channel selection for eeg-based depression detection via kernel-target alignment", *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2545–2556, 2020.

[31] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, "Feature-level fusion approaches based on multimodal eeg data for depression recognition", *Information Fusion*, vol. 59, pp. 127–138, 2020.

[32] Q. Chen, I. Chaturvedi, S. Ji, and E. Cambria, "Sequential fusion of facial appearance and dynamics for depression recognition", *Pattern Recognition Letters*, vol. 150, pp. 115–121, 2021.

[33] M. Tadalagi and A. M. Joshi, "Autodep: Automatic depression detection using facial expressions based on linear binary pattern descriptor", *Medical & biological engineering & computing*, vol. 59, no. 6, pp. 1339–1354, 2021.

[34] Z. Huang, J. Epps, and D. Joachim, "Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments", in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6549–6553.

[35] S. R. Müller, X. L. Chen, H. Peters, A. Chaintreau, and S. C. Matz, "Depression predictions from gps-based mobility do not generalize well to large demographically heterogeneous samples", *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[36] S. Ware, C. Yue, R. Morillo, *et al.*, "Large-scale automatic depression screening using meta-data from wifi infrastructure", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, pp. 1–27, 2018.

[37] P. Chikersal, A. Doryab, M. Tumminia, *et al.*, "Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection", *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 28, no. 1, pp. 1–41, 2021.

[38] J. C. Tolentino and S. L. Schmidt, "Dsm-5 criteria and depression severity: Implications for clinical practice", *Frontiers in psychiatry*, vol. 9, p. 450, 2018. DOI: `https://doi.org/10.3389/fpsyt.2018.00450`.

[39] G. Tutz, "Ordinal regression: A review and a taxonomy of models", *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 14, no. 2, e1545, 2022.

[40] B.-B. Gao, X.-X. Liu, H.-Y. Zhou, J. Wu, and X. Geng, "Learning expectation of label distribution for facial age and attractiveness estimation", *arXiv preprint arXiv:2007.01771*, 2020.

[41] H. Pan, H. Hu, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face", in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[42] C. Zhang, S. Liu, X. Xu, and C. Zhu, "C3ae: Exploring the limits of compact model for age estimation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 587–12 596.

[43] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang, "Ssr-net: A compact soft stagewise regression network for age estimation.", in *IJCAI*, vol. 5, 2018, p. 7.

[44] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. L. Yuille, "Deep regression forests for age estimation", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2304–2313, 2018.

[45] Y. Liu, F. Wang, and A. W. K. Kong, "Probabilistic deep ordinal regression based on gaussian processes", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5301–5309.

[46] Y. Zhou, J. He, and H. Gu, "Partial label learning via gaussian processes", *IEEE transactions on cybernetics*, vol. 47, no. 12, pp. 4443–4450, 2016.

[47] Y.-Y. Fan, S. Liu, B. Li, *et al.*, "Label distribution-based facial attractiveness computation by deep residual learning", *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2196–2208, 2017.

[48] X. Wen, B. Li, H. Guo, *et al.*, "Adaptive variance based label distribution learning for facial age estimation", in *ECCV*, 2020.

[49] A. Berg, M. Oskarsson, and M. O'Connor, "Deep ordinal regression with label diversity", in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 2740–2747.

[50] W. Li, X. Huang, J. Lu, J. Feng, and J. Zhou, "Learning probabilistic ordinal embeddings for uncertainty-aware regression", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 896–13 905.

[51] G. Chen, J. Peng, L. Wang, H. Yuan, and Y. Huang, "Feature constraint reinforcement based age estimation", *Multimedia Tools and Applications*, pp. 1–22, 2022.

[52] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models", *arXiv preprint arXiv:2106.09685*, 2021.

[53] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation", *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. abs/2101.00190, 2021. [Online]. Available: `https://api.semanticscholar.org/CorpusID:230433941`.

[54] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning", *arXiv preprint arXiv:2104.08691*, 2021.

[55] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning", *arXiv preprint arXiv:2110.04366*, 2021.

[56] N. Houlsby, A. Giurgiu, S. Jastrzebski, *et al.*, "Parameter-efficient transfer learning for nlp", in *International Conference on Machine Learning*, PMLR, 2019, pp. 2790–2799.

[57] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners", *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[58] J. Wei, X. Wang, D. Schuurmans, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models", *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[59] M. Parmar, S. Mishra, M. Purohit, M. Luo, M. H. Murad, and C. Baral, "In-boxbart: Get instructions into biomedical multi-task learning", *arXiv preprint arXiv:2204.07600*, 2022.

[60] Y. Wang, S. Mishra, P. Alipoormolabashi, *et al.*, "Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks", *arXiv:2204.07705*, 2022.

[61] L. Ouyang, J. Wu, X. Jiang, *et al.*, "Training language models to follow instructions with human feedback", *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[62] Y. Wang, Y. Kordi, S. Mishra, *et al.*, "Self-instruct: Aligning language model with self generated instructions", *arXiv preprint arXiv:2212.10560*, 2022.

[63] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine", *Nature Medicine*, pp. 1–11, 2023.

[64] L. Yang, H. Chen, Z. Li, X. Ding, and X. Wu, "Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-aware language modeling", *arXiv preprint arXiv:2306.11489*, 2023.

[65] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large language models in health care: Development, applications, and challenges", *Health Care Science*, 2023.

[66] O. Honovich, T. Scialom, O. Levy, and T. Schick, "Unnatural instructions: Tuning language models with (almost) no human labor", *arXiv preprint arXiv:2212.09689*, 2022.

[67] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models", *arXiv preprint arXiv:2302.13971*, 2023a.

[68] H. Touvron, L. Martin, K. Stone, *et al.*, "Llama 2: Open foundation and fine-tuned chat models", *arXiv preprint arXiv:2307.09288*, 2023b.

[69] Y. Wang, S. Mishra, P. Alipoormolabashi, *et al.*, "Benchmarking generalization via in-context instructions on 1,600+ language tasks", *arXiv e-prints*, arXiv–2204, 2022.

[70] B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with gpt-4", *arXiv preprint arXiv:2304.03277*, 2023.

[71] H. Mizuhara, L.-Q. Wang, K. Kobayashi, and Y. Yamaguchi, "Long-range eeg phase synchronization during an arithmetic task indexes a coherent cortical network simultaneously measured by fmri", *Neuroimage*, vol. 27, no. 3, pp. 553–563, 2005, ISSN: 1053-8119.

[72] H. Mizuhara and Y. Yamaguchi, "Human cortical circuits for central executive function emerge by theta phase synchronization", *Neuroimage*, vol. 36, no. 1, pp. 232–244, 2007, ISSN: 1053-8119.

[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

[74] J. Jeganathan, A. Perry, D. S. Ba Ssett, G. Roberts, P. B. Mitchell, and M. Breakspear, "Fronto-limbic dysconnectivity leads to impaired brain network controllability in young people with bipolar disorder and those at high genetic risk", *Neuroimage Clinical*, vol. 19, pp. 71–81, 2018.

[75] S. M. Strakowski, C. M. Adler, S. K. Holland, N. P. Mills, and J. C. Eliassen, "Abnormal fmri brain activation in euthymic bipolar disorder patients during a counting stroop interference task", *American Journal of Psychiatry*, vol. 162, no. 9, pp. 1697–1705, 2016.

[76] M. E. Lynall, D. S. Bassett, R. Kerwin, *et al.*, "Functional connectivity and brain networks in schizophrenia", *Journal of Neuroscience*, vol. 30, no. 28, pp. 9477–9487, 2010.

[77] H. Cai, J. Han, Y. Chen, *et al.*, "A pervasive approach to eeg-based depression detection", *Complexity*, vol. 2018, 2018, ISSN: 1076-2787.

[78] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, "Feature-level fusion approaches based on multimodal eeg data for depression recognition", *Information Fusion*, vol. 59, pp. 127–138, 2020, ISSN: 1566-2535.

[79] S. D. Puthankattil and P. K. Joseph, "Classification of eeg signals in normal and depression conditions by ann using rwe and signal entropy", *Journal of Mechanics in Medicine and biology*, vol. 12, no. 04, p. 1 240 019, 2012, ISSN: 0219-5194.

[80] S. Aydın, "Cross-validated adaboost classification of emotion regulation strategies identified by spectral coherence in resting-state", *Neuroinformatics*, pp. 1–13, 2021, ISSN: 1559-0089.

[81] A. F. Leuchter, I. A. Cook, A. M. Hunter, C. Cai, and S. Horvath, "Resting-state quantitative electroencephalography reveals increased neurophysiologic connectivity in depression", *PloS one*, vol. 7, no. 2, e32508, 2012, ISSN: 1932-6203.

[82] B. Li, K. Friston, M. Mody, H. Wang, H. Lu, and D. Hu, "A brain network model for depression: From symptom understanding to disease intervention", *CNS neuroscience & therapeutics*, vol. 24, no. 11, pp. 1004–1019, 2018, ISSN: 1755-5930.

[83] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277.

[84] P. Beloe and N. Derakshan, "Adaptive working memory training can reduce anxiety and depression vulnerability in adolescents", *Developmental science*, e12831, 2019, ISSN: 1363-755X.

[85] S. J. Bruijniks, G. van Grootheest, P. Cuijpers, *et al.*, "Working memory moderates the relation between the brain-derived neurotropic factor (bdnf) and psychotherapy outcome for depression", *Journal of Psychiatric Research*, vol. 130, pp. 424–432, 2020, ISSN: 0022-3956.

[86] B. Hosseinifard, M. H. Moradi, and R. Rostami, "Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from eeg signal", *Computer methods and programs in biomedicine*, vol. 109, no. 3, pp. 339–345, 2013, ISSN: 0169-2607.

[87] S.-C. Liao, C.-T. Wu, H.-C. Huang, W.-T. Cheng, and Y.-H. Liu, "Major depression detection from eeg signals using kernel eigen-filter-bank common spatial patterns", *Sensors*, vol. 17, no. 6, p. 1385, 2017.

[88] G. M. Bairy, O. S. Lih, Y. Hagiwara, *et al.*, "Automated diagnosis of depression electroencephalograph signals using linear prediction coding and higher order spectra features", *Journal of Medical Imaging and Health Informatics*, vol. 7, no. 8, pp. 1857–1862, 2017, ISSN: 2156-7018.

[89] K. Yoshida, Y. Shimizu, J. Yoshimoto, *et al.*, "Prediction of clinical depression scores and detection of changes in whole-brain using resting-state functional mri data with partial least squares regression", *PloS one*, vol. 12, no. 7, e0179638, 2017, ISSN: 1932-6203.

[90] M. Tanaka, Y. Shigihara, M. Funakura, E. Kanai, and Y. Watanabe, "Fatigue-associated alterations of cognitive function and electroencephalographic power densities", *PLoS One*, vol. 7, no. 4, e34774, 2012, ISSN: 1932-6203.

[91] A. Yassin, A.-H. Al-Mistarehi, K. El-Salem, *et al.*, "Clinical, radiological, and electroencephalographic features of hhv-6 encephalitis following hematopoietic stem cell transplantation", *Annals of Medicine and Surgery*, vol. 60, pp. 81–86, 2020, ISSN: 2049-0801.

[92] F. Zhang, F. Wang, C.-H. Li, *et al.*, "Therapeutic effects of subthalamic nucleus deep brain stimulation on anxiety and depression in parkinson's disease patients", *Journal of Neurorestoratology*, vol. 10, no. 1, pp. 31–42, 2022.

[93] B. K. Prusty, N. Gulve, S. Govind, *et al.*, "Active hhv-6 infection of cerebellar purkinje cells in mood disorders", *Frontiers in microbiology*, vol. 9, p. 1955, 2018, ISSN: 1664-302X.

[94] N. Kobayashi, N. Oka, M. Takahashi, *et al.*, "Human herpesvirus 6b greatly increases risk of depression by activating hypothalamic-pituitary-adrenal axis during latent phase of infection", *iScience*, p. 101 187, 2020, ISSN: 2589-0042.

[95] O. Murphy, K. Hoy, D Wong, N. Bailey, P. B. Fitzgerald, and R. Segrave, "Individuals with depression display abnormal modulation of neural oscillatory activity during working memory encoding and maintenance", *Biological psychology*, vol. 148, p. 107 766, 2019, ISSN: 0301-0511.

[96] A. A. Fingelkurts and A. A. Fingelkurts, "Altered structure of dynamic electroencephalogram oscillatory pattern in major depression", *Biological Psychiatry*, vol. 77, no. 12, pp. 1050–1060, 2015, ISSN: 0006-3223.

[97] A. A. Fingelkurts, A. A. Fingelkurts, H. Rytsälä, K. Suominen, E. Isometsä, and S. Kähkönen, "Composition of brain oscillations in ongoing eeg during major depression disorder", *Neuroscience research*, vol. 56, no. 2, pp. 133–144, 2006, ISSN: 0168-0102.

[98] Y. Pathak, O. Salami, S. Baillet, Z. Li, and C. R. Butson, "Longitudinal changes in depressive circuitry in response to neuromodulation therapy", *Frontiers in neural circuits*, vol. 10, p. 50, 2016, ISSN: 1662-5110.

[99] K. Benchenane, P. H. Tiesinga, and F. P. Battaglia, "Oscillations in the prefrontal cortex: A gateway to memory and attention", *Current opinion in neurobiology*, vol. 21, no. 3, pp. 475–485, 2011, ISSN: 0959-4388.

[100] X. Zhang, J. Li, K. Hou, B. Hu, J. Shen, and J. Pan, "Eeg-based depression detection using convolutional neural network with demographic attention mechanism", in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, pp. 128–133, ISBN: 1728119901.

[101] X. Li, R. La, Y. Wang, *et al.*, "Eeg-based mild depression recognition using convolutional neural network", *Medical & biological engineering & computing*, vol. 57, no. 6, pp. 1341–1352, 2019, ISSN: 1741-0444.

[102] R. L. Spitzer, *Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I), Clinician Version, User's Guide*. 1996.

[103] B. Mwangi, K. Matthews, and J. D. Steele, "Prediction of illness severity in patients with major depression using structural mr brain scans", *Journal of Magnetic Resonance Imaging*, vol. 35, no. 1, pp. 64–71, 2012, ISSN: 1053-1807.

[104] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity", *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.

[105] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep feature learning for facial age estimation", *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 157–164, 2017.

[106] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression", in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, 2006, pp. 341–345. DOI: `10.1109/FGR.2006.78`.

[107] G. Panis and A. Lanitis, "An overview of research activities in facial age estimation using the fg-net aging database", *Springer International Publishing*, 2014.

[108] B. C. Chen, C. S. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset", *IEEE Transactions on Multimedia*, vol. 17, no. 6, pp. 804–815, 2015.

[109] N.-H. Shin, S.-H. Lee, and C.-S. Kim, "Moving window regression: A novel approach to ordinal regression", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 760–18 769.

[110] S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures for ordinal regression", in *2009 Ninth international conference on intelligent systems design and applications*, IEEE, 2009, pp. 283–287.

[111] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 34–42.

[112] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image", in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 10–15.

[113] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4920–4928.

[114] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation", *Pattern Recognition Letters*, vol. 140, pp. 325–331, 2020.

[115] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks", *International Journal of Computer Vision*, vol. 126, no. 2, pp. 144–157, 2018.

[116] J. C. Gámez, D. Garcia, A. González, and R. Perez, "An approximation to solve regression problems with a genetic fuzzy rule ordinal algorithm", *Applied Soft Computing*, vol. 78, pp. 13–28, 2019.

[117] J. Alcalá-Fdez, R. Alcalá, S. González, Y. Nojima, and S. García, "Evolutionary fuzzy rule-based methods for monotonic classification", *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 6, pp. 1376–1390, 2017.

[118] R. Vega, P. Gorji, Z. Zhang, *et al.*, "Sample efficient learning of image-based diagnostic classifiers via probabilistic labels", in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2021, pp. 739–747.

[119] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[120] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network", *arXiv preprint arXiv:1503.02531*, 2015.

[121] X. Geng, "Label distribution learning", *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[122] E. Imani and M. White, "Improving regression performance with distributional losses", in *International Conference on Machine Learning*, PMLR, 2018, pp. 2157–2166.

[123] W. Chu, Z. Ghahramani, and C. K. Williams, "Gaussian processes for ordinal regression.", *Journal of machine learning research*, vol. 6, no. 7, 2005.

[124] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep feature learning for facial age estimation", in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, IEEE, 2017, pp. 157–164.

[125] Z. Zhang, C. Lai, H. Liu, and Y.-F. Li, "Infrared facial expression recognition via gaussian-based label distribution learning in the dark illumination environment for human emotion detection", *Neurocomputing*, vol. 409, pp. 341–350, 2020.

[126] G. P. Rajasekhar, E. Granger, and P. Cardinal, "Deep domain adaptation with ordinal regression for pain assessment using weakly-labeled videos", *Image and Vision Computing*, vol. 110, p. 104 167, 2021.

[127] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-cnn for age estimation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5183–5192.

[128] S. Chen, C. Zhang, and M. Dong, "Deep age estimation: From classification to ranking", *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2209–2222, 2017.

[129] K. Li, J. Xing, W. Hu, and S. J. Maybank, "D2c: Deep cumulatively and comparatively learning for human age estimation", *Pattern Recognition*, vol. 66, pp. 95–105, 2017.

[130] Z. Tan, S. Zhou, J. Wan, Z. Lei, and S. Z. Li, "Age estimation based on a single network with soft softmax of aging modeling", in *Asian Conference on Computer Vision*, Springer, 2016, pp. 203–216.

[131] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world", *ArXiv*, vol. abs/1607.02533, 2017.

[132] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses", *arXiv:1705.07204*, 2017.

[133] Y. Dong, F. Liao, T. Pang, *et al.*, "Boosting adversarial attacks with momentum", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[134] T. O. Hodson, "Root-mean-square error (rmse) or mean absolute error (mae): When to use them or not", *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, 2022.

[135] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance", *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[136] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature", *Geoscientific model development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[137] K. J. Geras, S. Wolfson, Y. Shen, *et al.*, "High-resolution breast cancer screening with multi-view deep convolutional neural networks", *arXiv:1703.07047*, 2017.

[138] A. Akselrod-Ballin, L. Karlinsky, S. Alpert, S. Hasoul, R. Ben-Ari, and E. Barkan, "A region based convolutional network for tumor detection and classification in breast mammography", in *Deep learning and data labeling for medical applications*, Springer, 2016, pp. 197–205.

[139] Y. Lin, J. Shen, Y. Wang, and M. Pantic, "Fp-age: Leveraging face parsing attention for facial age estimation in the wild", *ArXiv*, vol. abs/2106.11145, 2021.

[140] Z. Deng, H. Liu, Y. Wang, C. Wang, Z. Yu, and X. Sun, "Pml: Progressive margin loss for long-tailed age classification", *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 498–10 507, 2021.

[141] S. Hashempour, R. Boostani, M. Mohammadi, and S. Sanei, "Continuous scoring of depression from eeg signals via a hybrid of convolutional neural networks", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 176–183, 2022.

[142] C. Constantinidis and T. Klingberg, "The neuroscience of working memory capacity and training", *Nature Reviews Neuroscience*, vol. 17, no. 7, pp. 438–449, 2016.

[143] M. J. Kane, L. H. Brown, J. C. McVay, P. J. Silvia, I. Myin-Germeys, and T. R. Kwapil, "For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life", *Psychological science*, vol. 18, no. 7, pp. 614–621, 2007.

[144] S. E. Gathercole, L. Brown, and S. J. Pickering, "Working memory assessments at school entry as longitudinal predictors of national curriculum attainment levels", *Educational and Child Psychology*, vol. 20, no. 3, pp. 109–122, 2003.

[145] J. Eriksson, E. K. Vogel, A. Lansner, F. Bergström, and L. Nyberg, "Neurocognitive architecture of working memory", *Neuron*, vol. 88, no. 1, pp. 33–46, 2015.

[146] M. D'Esposito and B. R. Postle, "The cognitive neuroscience of working memory", *Annual review of psychology*, vol. 66, pp. 115–142, 2015.

[147] T. Pasternak and M. W. Greenlee, "Working memory in primate sensory systems", *Nature Reviews Neuroscience*, vol. 6, no. 2, pp. 97–107, 2005.

[148] R. A. Charlton, T. R. Barrick, I. N. C. Lawes, H. S. Markus, and R. G. Morris, "White matter pathways associated with working memory in normal aging", *Cortex*, vol. 46, no. 4, pp. 474–489, 2010.

[149] D. E. Nee, J. W. Brown, M. K. Askren, *et al.*, "A meta-analysis of executive components of working memory", *Cerebral cortex*, vol. 23, no. 2, pp. 264–282, 2013.

[150] N. Dolu, C Başar-Eroğlu, Ç Özesmi, and C Süer, "An assessment of working memory using p300 wave in healthy subjects", in *International Congress Series*, Elsevier, vol. 1278, 2005, pp. 7–10.

[151] A. M. Owen, K. M. McMillan, A. R. Laird, and E. Bullmore, "N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies", *Human brain mapping*, vol. 25, no. 1, pp. 46–59, 2005.

[152] T. D. Wager and E. E. Smith, "Neuroimaging studies of working memory", *Cognitive, Affective, & Behavioral Neuroscience*, vol. 3, pp. 255–274, 2003.

[153] K. Kubota and H. Niki, "Prefrontal cortical unit activity and delayed alternation performance in monkeys.", *Journal of neurophysiology*, vol. 34, no. 3, pp. 337–347, 1971.

[154] J. M. Fuster and G. E. Alexander, "Neuron activity related to short-term memory", *Science*, vol. 173, no. 3997, pp. 652–654, 1971.

[155] U. Leon-Dominguez, J. F. Martín-Rodríguez, and J. León-Carrión, "Executive n-back tasks for the neuropsychological assessment of working memory", *Behavioural brain research*, vol. 292, pp. 167–173, 2015.

[156] F. Collette, M. Hogge, E. Salmon, and M. Van der Linden, "Exploration of the neural substrates of executive functioning by functional neuroimaging", *Neuroscience*, vol. 139, no. 1, pp. 209–221, 2006.

[157] F. Collette, M. Van der Linden, S. Laureys, *et al.*, "Exploring the unity and diversity of the neural substrates of executive functioning", *Human brain mapping*, vol. 25, no. 4, pp. 409–423, 2005.

[158] M. Koenigs, A. K. Barbey, B. R. Postle, and J. Grafman, "Superior parietal cortex is critical for the manipulation of information in working memory", *Journal of Neuroscience*, vol. 29, no. 47, pp. 14 980–14 986, 2009.

[159] B. R. Buchsbaum and M. D'Esposito, "The search for the phonological store: From loop to convolution", *Journal of Cognitive Neuroscience*, vol. 20, no. 5, pp. 762–778, 2008.

[160] N. E. Myers, M. G. Stokes, and A. C. Nobre, "Prioritizing information during working memory: Beyond sustained internal attention", *Trends in cognitive sciences*, vol. 21, no. 6, pp. 449–461, 2017.

[161] A. Ikkai and C. E. Curtis, "Common neural mechanisms supporting spatial working memory, attention and motor intention", *Neuropsychologia*, vol. 49, no. 6, pp. 1428–1434, 2011.

[162] T. A. Jerde, E. P. Merriam, A. C. Riggall, J. H. Hedges, and C. E. Curtis, "Prioritized maps of space in human frontoparietal cortex", *Journal of Neuroscience*, vol. 32, no. 48, pp. 17 382–17 390, 2012.

[163] J. A. Cromer, J. E. Roy, T. J. Buschman, and E. K. Miller, "Comparison of primate prefrontal and premotor cortex neuronal activity during visual categorization", *Journal of cognitive neuroscience*, vol. 23, no. 11, pp. 3355–3365, 2011.

[164] J. E. Roy, M. Riesenhuber, T. Poggio, and E. K. Miller, "Prefrontal cortex activity during flexible categorization", *Journal of Neuroscience*, vol. 30, no. 25, pp. 8519–8528, 2010.

[165] M. G. Stokes, M. Kusunoki, N. Sigala, H. Nili, D. Gaffan, and J. Duncan, "Dynamic coding for cognitive control in prefrontal cortex", *Neuron*, vol. 78, no. 2, pp. 364–375, 2013.

[166] R. Quentin, J.-R. King, E. Sallard, *et al.*, "Differential brain mechanisms of selection and maintenance of information during working memory", *Journal of Neuroscience*, vol. 39, no. 19, pp. 3728–3740, 2019.

[167] J. M. Fuster and S. L. Bressler, "Cognit activation: A mechanism enabling temporal integration in working memory", *Trends in cognitive sciences*, vol. 16, no. 4, pp. 207–218, 2012.

[168] G. S. Shields, J. C. Bonner, and W. G. Moons, "Does cortisol influence core executive functions? a meta-analysis of acute cortisol administration effects on working memory, inhibition, and set-shifting", *Psychoneuroendocrinology*, vol. 58, pp. 91–103, 2015.

[169] C. Rottschy, R. Langner, I. Dogan, *et al.*, "Modelling neural correlates of working memory: A coordinate-based meta-analysis", *Neuroimage*, vol. 60, no. 1, pp. 830–846, 2012.

[170] L. Hu, Z. Zhang, and Y. Hu, "A time-varying source connectivity approach to reveal human somatosensory information processing", *Neuroimage*, vol. 62, no. 1, pp. 217–228, 2012.

[171] F. Li, B. Chen, H. Li, *et al.*, "The time-varying networks in p300: A task-evoked eeg study", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 7, pp. 725–733, 2016.

[172] G Tropini, J Chiang, Z. Wang, and M. McKeown, "Partial directed coherence-based information flow in parkinson's disease patients performing a visually-guided motor task", in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2009, pp. 1873–1878.

[173] H. Mizuhara and Y. Yamaguchi, "Human cortical circuits for central executive function emerge by theta phase synchronization", *NeuroImage*, vol. 36, no. 1, pp. 232–244, 2007, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2007.02.026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811907001085.

[174] H. Mizuhara, L.-Q. Wang, K. Kobayashi, and Y. Yamaguchi, "Long-range eeg phase synchronization during an arithmetic task indexes a coherent cortical network simultaneously measured by fmri", *NeuroImage*, vol. 27, no. 3, pp. 553–563, 2005, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2005.04.030. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811905002788.

[175] L. A. Baccalá and K. Sameshima, "Partial directed coherence: A new concept in neural structure determination", *Biological cybernetics*, vol. 84, no. 6, pp. 463–474, 2001.

[176] L. A. Baccala, K. Sameshima, and D. Y. Takahashi, "Generalized partial directed coherence", in *2007 15th International conference on digital signal processing*, Ieee, 2007, pp. 163–166.

[177] B. Scott L and M. Earl K, "Frequency-specific hippocampalprefrontal interactions during associative learning", vol. 18, Springer, 2015, pp. 576–581.

[178] J. Dauwels, F. B. Vialatte, T. Musha, and A. Cichocki, "A comparative study of synchrony measures for the early diagnosis of alzheimer's disease based on eeg", *NeuroImage*, vol. 49, pp. 668–693, 2010. [Online]. Available: `https://api.semanticscholar.org/CorpusID:5708896`.

[179] J. Taxidis, B. Coomber, R. Mason, and M. R. Owen, "Assessing cortico-hippocampal functional connectivity under anesthesia and kainic acid using generalized partial directed coherence", *Biological Cybernetics*, vol. 102, pp. 327–340, 2010. [Online]. Available: `https://api.semanticscholar.org/CorpusID:1173721`.

[180] P.-O. Harvey, P. Fossati, J.-B. Pochon, *et al.*, "Cognitive control and brain resources in major depression: An fmri study using the n-back task", *NeuroImage*, vol. 26, pp. 860–869, 2005. [Online]. Available: `https://api.semanticscholar.org/CorpusID:39617212`.

[181] M. Scherg, "Functional imaging and localization of electromagnetic brain activity", *Brain Topography*, vol. 5, pp. 103–111, 2005. [Online]. Available: `https://api.semanticscholar.org/CorpusID:21499919`.

[182] M. Scherg and P. Berg, "New concepts of brain source imaging and localization.", *Electroencephalography and clinical neurophysiology. Supplement*, vol. 46, pp. 127–37, 1996. [Online]. Available: `https://api.semanticscholar.org/CorpusID:46405309`.

[183] M. P. Tarvainen, J. K. Hiltunen, P. O. Ranta-aho, and P. A. Karjalainen, "Estimation of nonstationary eeg with kalman smoother approach: An application to event-related synchronization (ers)", *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 516–524, 2004. [Online]. Available: `https://api.semanticscholar.org/CorpusID:5423390`.

[184] J. Polich, "Updating p300: An integrative theory of p3a and p3b", *Clinical Neurophysiology*, vol. 118, pp. 2128–2148, 2007. [Online]. Available: `https://api.semanticscholar.org/CorpusID:9753972`.

[185] C. J. Stoodley and J. D. Schmahmann, "Functional topography in the human cerebellum: A meta-analysis of neuroimaging studies", *NeuroImage*, vol. 44, pp. 489–501, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:2703288.

[186] J. J. LaRocque, J. A. Lewis-Peacock, A. T. Drysdale, K. Oberauer, and B. R. Postle, "Decoding attended information in short-term memory: An eeg study", *Journal of Cognitive Neuroscience*, vol. 25, pp. 127–142, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:1071449.

[187] O. Barak and M. Tsodyks, "Working models of working memory", *Current Opinion in Neurobiology*, vol. 25, pp. 20–24, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:31958359.

[188] R. C. O'Reilly, "Biologically based computational models of high-level cognition", *Science*, vol. 314, pp. 91 –94, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:11620257.

[189] A. K. Engel and P. Fries, "Beta-band oscillations—signalling the status quo?", *Current Opinion in Neurobiology*, vol. 20, pp. 156–165, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:79336156.

[190] M. M. Shafi, Y. Zhou, J. Quintana, C. C. Chow, J. M. Fuster, and M. Bodner, "Variability in neuronal activity in primate cortex during working memory tasks", *Neuroscience*, vol. 146, pp. 1082–1108, 2007. [Online]. Available: https://api.semanticscholar.org/CorpusID:16256426.

[191] A. C. Riggall and B. R. Postle, "The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging", *The Journal of Neuroscience*, vol. 32, pp. 12990 –12 998, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:18395926.

[192] B. H. Silverstein, M. D. Snodgrass, H. Shevrin, and R. K. Kushwaha, "P3b, consciousness, and complex unconscious processing", *Cortex*, vol. 73, pp. 216–227, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206985968.

[193] W.-J. Huang, W.-W. Chen, and X Zhang, "The neurophysiology of p 300–an integrated review.", *European review for medical and pharmacological sciences*, vol. 19 8, pp. 1480–8, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:1948290.

[194] L. Nyberg, M. Andersson, K. Kauppi, *et al.*, "Age-related and genetic modulation of frontal cortex efficiency", *Journal of Cognitive Neuroscience*, vol. 26, pp. 746–754, 2014. [Online]. Available: `https://api.semanticscholar.org/CorpusID:13583069`.

[195] A. Compte, N. J.-B. Brunel, P. S. Goldman-Rakic, and X.-J. Wang, "Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model.", *Cerebral cortex*, vol. 10 9, pp. 910–23, 2000. [Online]. Available: `https://api.semanticscholar.org/CorpusID:7239548`.

[196] M. Wang, Y. Yang, C.-J. Wang, *et al.*, "Nmda receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex", *Neuron*, vol. 77, pp. 736–749, 2013. [Online]. Available: `https://api.semanticscholar.org/CorpusID:5659560`.

[197] J. D. Murray, A. Anticevic, M. Gancsos, *et al.*, "Linking microcircuit dysfunction to cognitive impairment: Effects of disinhibition associated with schizophrenia in a cortical working memory model.", *Cerebral cortex*, vol. 24 4, pp. 859–72, 2014. [Online]. Available: `https://api.semanticscholar.org/CorpusID:2630534`.

[198] P. M. Bays, "Spikes not slots: Noise in neural populations limits working memory", *Trends in Cognitive Sciences*, vol. 19, pp. 431–438, 2015. [Online]. Available: `https://api.semanticscholar.org/CorpusID:698568`.

[199] M. Starc, J. D. Murray, N. Santamauro, *et al.*, "Schizophrenia is associated with a pattern of spatial working memory deficits consistent with cortical disinhibition", *Schizophrenia Research*, vol. 181, pp. 107–116, 2017. [Online]. Available: `https://api.semanticscholar.org/CorpusID:3728957`.

[200] S. L. Franconeri, G. A. Alvarez, and P. Cavanagh, "Flexible cognitive resources: Competitive content maps for attention and memory", *Trends in Cognitive Sciences*, vol. 17, pp. 134–141, 2013. [Online]. Available: `https://api.semanticscholar.org/CorpusID:14596389`.

[201] J. L. Lobo, J. Del Ser, A. Bifet, and N. Kasabov, "Spiking neural networks and online learning: An overview and perspectives", *Neural Networks*, vol. 121, pp. 88–100, 2020.

[202] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context", *arXiv:1901.02860*, 2019.

[203] L. J. Borg-Graham, C. Monier, and Y. Fregnac, "Visual input evokes transient and strong shunting inhibition in visual cortical neurons", *Nature*, vol. 393, no. 6683, pp. 369–373, 1998.

[204] W. Huang, Y. Ke, J. Zhu, *et al.*, "Tresk channel contributes to depolarization-induced shunting inhibition and modulates epileptic seizures", *Cell Reports*, vol. 36, no. 3, p. 109 404, 2021.

[205] A. F. Agarap, "Deep learning using rectified linear units (relu)", *arXiv preprint arXiv:1803.08375*, 2018.

[206] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[207] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)", *arXiv preprint arXiv:1511.07289*, 2015.

[208] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions", *arXiv preprint arXiv:1710.05941*, 2017.

[209] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units", 2016.

[210] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks", *Advances in neural information processing systems*, vol. 30, 2017.

[211] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need", *Advances in neural information processing systems*, vol. 30, 2017.

[212] J. Kaplan, S. McCandlish, T. Henighan, *et al.*, "Scaling laws for neural language models", *arXiv preprint arXiv:2001.08361*, 2020.

[213] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv:1810.04805*, 2018.

[214] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners", *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[215] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer", *arXiv preprint arXiv:1910.10683*, 2019.

[216] S. Smith, M. Patwary, B. Norick, *et al.*, "Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model", *arXiv preprint arXiv:2201.11990*, 2022.

[217] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach", *arXiv preprint arXiv:1907.11692*, 2019.

[218] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations", *arXiv preprint arXiv:1909.11942*, 2019.

[219] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention", *arXiv preprint arXiv:2006.03654*, 2020.

[220] P. He, J. Gao, and W. Chen, "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing", *arXiv:2111.09543*, 2021.

[221] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity", *arXiv preprint arXiv:2101.03961*, 2021.

[222] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[223] Z. Liu, H. Hu, Y. Lin, *et al.*, "Swin transformer v2: Scaling up capacity and resolution", *arXiv preprint arXiv:2111.09883*, 2021.

[224] Y. Zhang, S. Sun, M. Galley, *et al.*, "Dialogpt: Large-scale generative pre-training for conversational response generation", *arXiv preprint arXiv:1911.00536*, 2019.

[225] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "Adapterfusion: Non-destructive task composition for transfer learning", *arXiv preprint arXiv:2005.00247*, 2020.

[226] E. B. Zaken, S. Ravfogel, and Y. Goldberg, "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models", *arXiv:2106.10199*, 2021.

[227] D. Guo, A. M. Rush, and Y. Kim, "Parameter-efficient transfer learning with diff pruning", *arXiv preprint arXiv:2012.07463*, 2020.

[228] A. Chavan, Z. Liu, D. Gupta, E. Xing, and Z. Shen, "One-for-all: Generalized lora for parameter-efficient fine-tuning", *arXiv preprint arXiv:2306.07967*, 2023.

[229] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient fine-tuning of quantized llms", *arXiv preprint arXiv:2305.14314*, 2023.

[230] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: Vgg and residual architectures", *Frontiers in neuroscience*, vol. 13, p. 95, 2019.

[231] N. Rathi and K. Roy, "Diet-snn: A low-latency spiking neural network with direct input encoding and leakage and threshold optimization", *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[232] C. Lee, S. S. Sarwar, P. Panda, G. Srinivasan, and K. Roy, "Enabling spike-based backpropagation for training deep neural network architectures", *Frontiers in neuroscience*, p. 119, 2020.

[233] B. Rueckauer, I.-A. Lungu, Y. Hu, M. Pfeiffer, and S.-C. Liu, "Conversion of continuous-valued deep networks to efficient event-driven networks for image classification", *Frontiers in neuroscience*, vol. 11, p. 682, 2017.

[234] W. Fang, Z. Yu, Y. Chen, T. Masquelier, T. Huang, and Y. Tian, "Incorporating learnable membrane time constant to enhance learning of spiking neural networks", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2661–2671.

[235] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

[236] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding", *arXiv preprint arXiv:1804.07461*, 2018.

[237] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text", *arXiv preprint arXiv:1606.05250*, 2016.

[238] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "Swag: A large-scale adversarial dataset for grounded commonsense inference", *arXiv preprint arXiv:1808.05326*, 2018.

[239] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans", *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[240] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[241] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad", *arXiv preprint arXiv:1806.03822*, 2018.

[242] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps", in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 9204–9215. [Online]. Available: `https://proceedings.neurips.cc/paper/2021/file/4cc05b35c2f937c5bd9e7d41d3686fff-Paper.pdf`.

[243] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units", *arXiv preprint arXiv:1508.07909*, 2015.

[244] Y. Zhou, A. I. Muresanu, Z. Han, *et al.*, "Large language models are human-level prompt engineers", *arXiv preprint arXiv:2211.01910*, 2022.

[245] L. C. Magister, J. Mallinson, J. Adamek, E. Malmi, and A. Severyn, "Teaching small language models to reason", *arXiv preprint arXiv:2212.08410*, 2022.

[246] R. Taori, I. Gulrajani, T. Zhang, *et al.*, *Stanford alpaca: An instruction-following llama model*, 2023.

[247] W.-L. Chiang, Z. Li, Z. Lin, *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality", *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[248] ShareGPT. "Sharegpt, https://sharegpt.com/". (2023), [Online]. Available: `https://sharegpt.com/`.

[249] K. Harrigian, C. Aguirre, and M. Dredze, "On the state of social media data for mental health research", *arXiv preprint arXiv:2011.05233*, 2020.

[250] A. Das, S. Selek, A. R. Warner, *et al.*, "Conversational bots for psychotherapy: A study of generative transformer models using domain-specific dialogues", in *Proceedings of the 21st Workshop on Biomedical Language Processing*, 2022, pp. 285–297.

[251] V. Pérez-Rosas, X. Sun, C. Li, Y. Wang, K. Resnicow, and R. Mihalcea, "Analyzing the quality of counseling conversations: The tell-tale signs of high-quality counseling", in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[252] M. R. Pacheco-Lorenzo, S. M. Valladares-Rodríguez, L. E. Anido-Rifón, and M. J. Fernández-Iglesias, "Smart conversational agents for the detection of neuropsychiatric disorders: A systematic review", *Journal of Biomedical Informatics*, vol. 113, p. 103 632, 2021.

[253] K. T. Pham, A. Nabizadeh, and S. Selek, "Artificial intelligence and chatbots in psychiatry", *Psychiatric Quarterly*, vol. 93, no. 1, pp. 249–253, 2022.

[254] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews", 2019.

[255] Y.-T. Tsai and W.-A. Lin, "Design of an intelligent cognition assistant for people with cognitive impairment", in *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, IEEE, 2018, pp. 1207–1212.

[256] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and conversational agents in mental health: A review of the psychiatric landscape", *The Canadian Journal of Psychiatry*, vol. 64, no. 7, pp. 456–464, 2019.

[257] O. Honovich, U. Shaham, S. R. Bowman, and O. Levy, "Instruction induction: From few examples to natural language task descriptions", *arXiv:2205.10782*, 2022.

[258] S. Ye, D. Kim, J. Jang, J. Shin, and M. Seo, "Guess the instruction! flipped learning makes language models stronger zero-shot learners", in *The Eleventh International Conference on Learning Representations*, 2022.

[259] T. Schick and H. Schütze, "Generating datasets with pretrained language models", *arXiv preprint arXiv:2104.07540*, 2021.

[260] V. Sanh, A. Webson, C. Raffel, *et al.*, "Multitask prompted training enables zero-shot task generalization", *arXiv preprint arXiv:2110.08207*, 2021.

[261] OpenAI. "Chatgpt, https://openai.com/blog/chatgpt/". (2023), [Online]. Available: `https://openai.com/blog/chatgpt/`.

[262] G. Penedo, Q. Malartic, D. Hesslow, *et al.*, "The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only", *arXiv preprint arXiv:2306.01116*, 2023. arXiv: `2306.01116`. [Online]. Available: `https://arxiv.org/abs/2306.01116`.

[263] Z. Du, Y. Qian, X. Liu, *et al.*, "Glm: General language model pretraining with autoregressive blank infilling", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 320–335.

[264] M. N. Team. "Introducing mpt-7b: A new standard for open-source, commercially usable llms". Accessed: 2023-05-05. (2023), [Online]. Available: `www.mosaicml.com/blog/mpt-7b` (visited on 05/05/2023).

[265] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries", in *Text summarization branches out*, 2004, pp. 74–81.

[266] A. Efrat, O. Honovich, and O. Levy, "Lmentry: A language model benchmark of elementary language tasks", *arXiv preprint arXiv:2211.02069*, 2022.

[267] S. Gururangan, A. Marasović, S. Swayamdipta, *et al.*, "Don't stop pretraining: Adapt language models to domains and tasks", *arXiv preprint arXiv:2004.10964*, 2020. DOI: `https://doi.org/10.48550/arXiv.2004.10964`.

[268] H. W. Chung, L. Hou, S. Longpre, *et al.*, "Scaling instruction-finetuned language models", *arXiv preprint arXiv:2210.11416*, 2022.