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, December 2023

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

Copyright © December 2023 Cheng Kang

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, December 2023

....
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 Perception (MLP)s and inhibition mechanisms. In the fifth project, I and my co-authors developed a psychotherapy chatbot fine-tuned on Large Language Model (LLM) 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

(Automaticly translated by Google Translate)

Detekce velké depresivní poruchy (MDD) těžila z pokročilých neuro- výpočetní metody a tradiční techniky strojového učení. Kromě toho nová technologie nické nástroje se snaží zmírnit utrpení pacientů. Obecně platí, že studie o míra detekce deprese je většinou příliš nízká na to, aby byla přenesena do klinických aplikací, a techniky týkající se psychologických terapií nebo pomoci silně spoléhají na konkrétní místa a časy. V této práci popisuji výsledky tří projektů, které se zabývají chal- Díky tomu je míra detekce deprese stabilnější s vyšší mírou přesnosti a vytvoření dostupného psychoterapeutického chatbota bez předběžného školení na obrovský jazyk datové sady, ale se silnějším výkonem. V druhé kapitole já a moji spoluautoři vyvinula systém Brain-Computer Interface (BCI) pro zpracování elektroencefalogramu (EEG) signály a vytváření dynamických funkčních mozkových sítí mezi depresí pacientů a zdravých kontrol. Mezitím dvě zbytkové neuronové sítě založené na vybrané EEG kanály a frekvence byly použity k detekci deprese ze zdraví a vyhodnotit závažnost deprese pomocí skóre strukturovaného klinického rozhovoru pro Poruchy osy I DSM-IV, verze pro klinického lékaře (SCID-CV). Ve třetí kapitole Já a můj spoluautoři navrhli román Fuzzy Window with the Gaussian Processed Labels (FW-GPL) metoda pro ordinální skórovací úlohy. S využitím procesu okna má tento model výhoda zpracování ordinálních dat, jako jsou lékařské snímky a EEG pacientů s různou intenzitou deprese. Ve čtvrté kapitole rozvíjet pokročilý výcvik resp dolaďování metod založených na znalostech neurověd, které jsem se svými spoluautory studoval funkční dynamiku mozku během pracovní paměti (WM) a zjistili jsme údržbu, inhibice a disinhibice by měly spolupracovat na zpracování informací v našem mozku. V závislosti na těchto zjištěních ve třetí kapitole jsme navrhli architekturu inspirovanou neurovědami. model, inhibice posunu v kapitole čtyři a výsledky této nové architektury na doladění downstream jazykových úloh prokázat efektivitu hradlování Multilayer Percepce (MLP) a inhibiční mechanismy. V pátém projektu já a moji spoluautoři vyvinul psychoterapeutického chatbota vyladěného na velkém jazykovém modelu (LLM). AlexanderStreet terapeutická a poradenská data a poskytla profesionálnější a kom- mon využil psychoterapeutických znalostí. Kromě přispívání vědeckých závěrů o V každém systému budou tyto metody sloužit také jako praktický rámec pro budoucí úsilí řešit problémy při detekci deprese a pomoci při 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

Abstract **iv**

Abstrakt **v**

Acknowledgements **vi**

List of Tables **xi**

List of Figures **xiv**

1 Introduction **1**
- 1.1 Goals of the Thesis 1
- 1.2 Background 2
  - 1.2.1 Detecting Depression 3
  - 1.2.2 Psychotherapy Using Large Language Models 6
- 1.3 Thesis Outline 9

2 Classifying and Scoring Major Depressive Disorders by Selecting Frequencies and Channels with the Use of Residual Neural Networks **12**
- 2.1 Introduction 13
- 2.2 Related Works 14
  - 2.2.1 Brain regions and extraction of functional networks 15
  - 2.2.2 Artificial neural networks utilization 16
- 2.3 Methodology 17
  - 2.3.1 Participants and EEGs Recording 17
  - 2.3.2 Working Memory Experiments 18
  - 2.3.3 Preprocessing of EEGs before Training 18
  - 2.3.4 Residual neural networks 19
- 2.4 Result 20
  - 2.4.1 Memory load comparison of behavioural results 20
  - 2.4.2 The Connections comparison 20
  - 2.4.3 Clusters between these Three Groups 22
  - 2.4.4 The result of classifying and scoring MDD patients 22
- 2.5 Discussion 23
  - 2.5.1 Possible inducing reason for getting depression 24
  - 2.5.2 Topological analysis 24
  - 2.5.3 Contribution of frequency and topological selection for classifying and scoring depressive patients 25
  - 2.5.4 State of the art for classifying depressive patients 26

2.5.5 State of the art for scoring depressive severities 26
2.6 Conclusion and future work 27

3 Fuzzy Windows with Gaussian Processed Labels for Ordinal Scoring Tasks 30
3.1 Introduction 30
3.2 Related Work 32
3.2.1 Ordinal Classification 33
3.2.2 Windows for Ordinal Classification 34
3.2.3 Fuzzy Scoring for Ordinal Classification 34
3.2.4 Soft Labels and Gaussian Processes 36
3.3 Our Method 37
3.3.1 Normalized Gaussian Processed Labels 37
3.3.2 Fuzzy Windows with Normalized Gaussian Processed Labels 39
3.4 Experiments 41
3.4.1 Datasets 42
3.4.2 Evaluation Metrics 43
3.4.3 Experiment Settings 44
3.4.4 Hardware and Software 44
3.5 Results and Analysis 44
3.5.1 Scoring Breast Cancer Images 45
3.5.2 Scoring Facial-Age Images 45
3.5.3 Scoring Depressive Severity using EEGs 46
3.6 Ablation and Discussion 47
3.6.1 Ablation Study I (Influence of the Number of Neurons) 47
3.6.2 Ablation Study II (Influence of the Length of the Window L_{Win}) . . 47
3.6.3 Ablation Study III (Incomplete Ordinal Image Data) 48
3.6.4 Advantage and Limitation 49
3.7 Conclusions 49

4 Brain Networks of Maintenance, Inhibition and Disinhibition During Working Memory 51
4.1 Introduction 52
4.2 Related Work 53
4.2.1 Pathway for Attention Arousal and Executive Function 53
4.2.2 Pathway for Coding and Decoding 53
4.2.3 Pathway for Sustained Brain Activity 54
4.2.4 Pathway for Lateral Inhibition 54
4.3 Methods 55
4.3.1 Participants 55
4.3.2 Experimental Procedures 55
4.3.3 EEG Recording 56
4.3.4 Data Analysis 56
4.4 Study Results 61
4.4.1 Behavioral Result 61
4.4.2 Scalp Topography Performance 61
4.4.3 Band-Specific Synchrony Reflects 61
4.4.4 Band-Specific Directionality Reflects 63
4.4.5 The Neurocognitive Architecture With Component Processes of WM 63

4.5 Discussion 65
4.5.1 The Maintenance Loop During WM 66
4.5.2 The Inhibition Loop During WM 68
4.5.3 Conclusion And Future Directions 68

5 InA: Inhibition Adaption On Pre-trained Language Models 70
5.1 Introduction 70
5.2 Problem Statement 72
5.3 Explanation of Shunting Inhibition 73
5.3.1 Shunting Inhibition (Gate with Inhibition) 73
5.3.2 Membrane Potentials and Threshold 74
5.4 Related Work 74
5.4.1 Transformer-based language models 74
5.4.2 Fine-tuning on NLP downstream tasks 75
5.4.3 Parameter-Efficient Fine-Tuning 76
5.4.4 Threshold and Inhibition 77
5.5 Inhibition Adaption 77
5.5.1 Inhibited Adaption 77
5.5.2 Inserting InA into Transformer 78
5.6 Experiments 79
5.6.1 Experiment Settings 79
5.6.2 Evaluation Datasets 80
5.6.3 Fine-Tuning Implementation Details 80
5.6.4 Results 80
5.6.5 Effectiveness: InA on Fine-tuning 81
5.6.6 InA on the Text Classification Task. 82
5.7 Analysis and Discussion 85
5.7.1 Difference Between LoRA and InA 86
5.7.2 Should we need inhibition during fine-tuning? And how does it work? 87
5.7.3 How to choose the inhibition level Inh_p and select a good rank s in real cases? 88
5.7.4 Can InA really inhibit irrelevant knowledge? How can InA inhibit them? 90
5.8 Conclusion 91

6 Domain Specific Assistant Instruction on Psychotherapy Chatbot 92
6.1 Introduction 92
6.2 Problem Statement 94
6.3 Related Work 95
6.3.1 Psychotherapy-based Conversational Systems 95
6.3.2 Instruction Data for Language 95
6.3.3 Parameter-Efficient Fine-Tuning Pre-trained Language Models . . . 96
6.4 Methodology 97
6.4.1 Data Collection 97
6.4.2 Prompting Templates for Task Identification 98
6.4.3 Assistant-Instruction 99
6.4.4 Generate and Expand Psychotherapy Instructions 99
6.5 Experiments 100
6.5.1 Experiments Settings 101

6.5.2 Tuning on Psychotherapy Data 101
6.5.3 Evaluation 103
6.6 Results 105
6.6.1 Performance on Revision 105
6.6.2 Generation on Psychotherapy Domain 106
6.6.3 Evaluation of Psychologists 106
6.6.4 Human Evaluation Agreement 107
6.7 Analysis and Discussion 107
6.8 Limitations 109
6.9 Conclusion 109
6.10 Appendix 109

7 Conclusion 119
7.1 Summary 119
7.2 Contributions and Achievements 120
7.3 Future Work 121

List of Candidate's Publications Related to the Thesis 122
7.4 Publications in Impacted Journals 122
7.5 Other Publications 123

References 149

List of Tables

2.1 The comparison of reaction time and response accuracy rates between two different memory loads (average ± standard deviation) in two depressive groups. 20
2.2 Classification (Accuracy) and scoring depression (Root-Mean-Square Error (RMSE)) results (mean and standard deviation) using whole frequency bands (delta, theta, alpha and beta). 22
2.3 Classification (Accuracy) and scoring depression (RMSE) results (mean and standard deviation) using beta frequency bands. 22
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. 22
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. 23
2.6 Comparison with existing methods on classifying depression with EEGs. . 26
2.7 Comparison with existing methods on scoring depressive severities with EEGs. 26

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. 38
3.2 Sample distribution of CBIS-DDSM dataset based on BI-RADS assessment. 42
3.3 Facial-age datasets used to evaluate the proposed FW-GPL. 43
3.4 Comparison with existing methods on DDSM in terms of Accuracy (ACC). 45
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.) 45
3.6 Comparison with existing methods on scoring depressive severities with EEGs. 46
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]). 47
3.8 Test performance of the DEX method (set length of output neurons N as [100, 50, 20, 10, 5]). 47
3.9 Test performance of FW-GPL on testing data sets (length of output neurons set as 100). 48
3.10 Test performance of FW-GPL on testing data sets (length of output neurons set as 10). 48

3.11 Test performance between the DEX and the FW-GPL with fragmentary IMDB-WIKI dataset. (length of output neurons set as 100). 49
3.12 Test performance between the DEX and the FW-GPL with fragmentary IMDB-WIKI dataset. (length of the window set as 10). 49

5.1 Hyper-parameters for fine-tuning Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized BERT (RoBERTa) and Decoding-enhanced BERT with Disentangled Attention (DeBERTa) with inhibited gate MLPs mechanism on down-streaming tasks. 79
5.2 The efficiency of Inhibition Adaption (InA) and other adaptation fine-tuning (FT) methods in terms of trainable parameters, update speed (back-propagation) and inference (complexity). 81
5.3 Comparison results of fine-tuning the General Language Understanding Evaluation (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 [55] for a fair comparison. 81
5.4 Comparison results of fine-tuning Stanford Question Answering Dataset (SQuAD) v1.1, SQuAD v2.0 and Situations With Adversarial Generations (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 '-'). 82
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 $BERT - large$. 84
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. 84
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). 85
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. 86

6.1 The Natural-Instruction and GPT-4 revised Assistant-Instruction on the Depressive Disorder domain. 97
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. 98
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. 100
6.4 Hyper-parameters for querying OpenAI API in different experiments. . . . 101
6.5 Hyper-parameters for fine-tuning pre-trained LLMs in different experiments. 102
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). 103

6.7 The Natural-Instruction and GPT-4 revised Assistant-Instruction on the Bipolar Disorders domain. 104
6.8 The Natural-Instruction and GPT-4 revised Assistant-Instruction on the Addiction domain. 105
6.9 The Natural-Instruction and GPT-4 revised Assistant-Instruction on the Addictive Disorder domain. 111
6.10 The Zero-Shot, Natural-Instruction and Assistant-Instruction Tuned ChatGLM-6B on Psychotherapy data. 112
6.11 The Zero-Shot, Natural-Instruction and Assistant-Instruction Tuned MPT-7B on Psychotherapy data. 113
6.12 The Zero-Shot, Natural-Instruction and Assistant-Instruction Tuned Falcon-7B on Psychotherapy data. 114
6.13 The Zero-Shot, Natural-Instruction and Assistant-Instruction Tuned Llama2-7B on Psychotherapy data. 115
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. 116
6.15 One example of using Llama2-7B. $M1$, $M2$ and $M3$ respectively is the Readability, Professional and Match Score (original LLM / finetuned-LLM). 116
6.16 One example of using ChatGLM2-6B. $M1$, $M2$ and $M3$ respectively is the Readability, Professional and Match Score (original LLM / finetuned-LLM). 117
6.17 Two examples of using MPT-7B. $M1$, $M2$ and $M3$ respectively is the Readability, Professional and Match Score (original LLM / finetuned-LLM). . . 118

List of Figures

1.1 The whole framework of depression detection and assistance in this thesis. 2
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**). 4
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. 5
1.4 A practical example of InA is its use in the $BERT_{large}$ model, which has been fine-tuned under question-answering datasets. 6
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. 8

2.1 The structure of the constructed residual neural network. The input size is $64 \times 64 \times 18$ or $16 \times 64 \times 18$. 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. ×3 means this block should be repeated triple times. 20
2.2 The number of the significant pairs in terms of the comparison between 2-back and 0-back tasks. 21
2.3 The t values (significant level) of the comparison between the depression group and the healthy control group. 28
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. 29

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. 33

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. 35

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 $C_{i\text{-}1}$. 40

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. 48

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). 55

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 Global Field Power (GFP) (the sum of squares of all channels, normalized to 100%) shown in the central is displayed in a logarithmic scale.. 57

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. 58

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 ∼ 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. 62

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. 64

4.6 Schematic explanation of representations to brain networks during WM tasks. Left upper panel is the location illustration of four fitted sources. **A∼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. 65

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. 66

5.1 Illustration of the transformer architecture and our proposed parameter-efficient tuning method: inhibition adaption. 72

5.2 Inspiration from Neuroscience: Gate With Inhibition. 75

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. . . . 82

5.4 Roughly disassembled DeBERTaV3 architecture. 86

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). 87

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). 87

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). 88
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). 88
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). 89
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). 89

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. 93
6.2 The schematic of the model fine-tuning and conversation between Chatbot and User. 96

Chapter 1

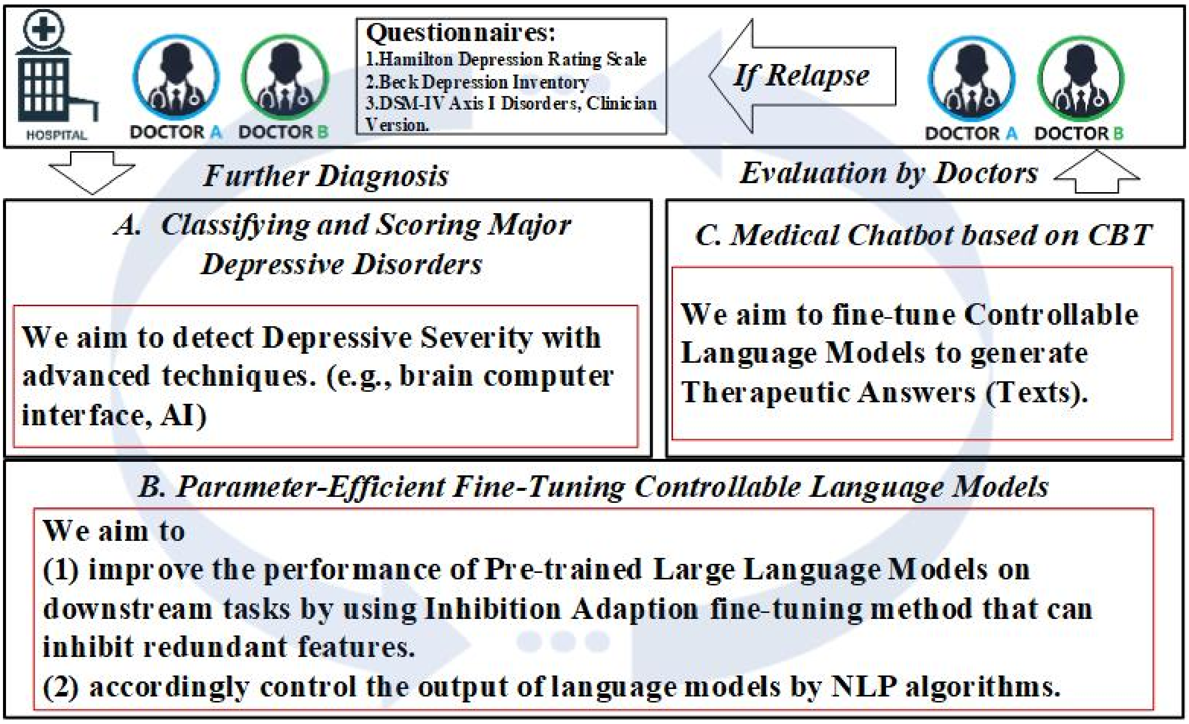
Introduction

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 Network (ANN)s, 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.


HOSPITAL
DOCTOR A
DOCTOR B
Questionnaires:
1.Hamilton Depression Rating Scale
2.Beck Depression Inventory
3.DSM-IV Axis I Disorders, Clinician Version.
If Relapse
DOCTOR A
DOCTOR B
Further Diagnosis
Evaluation by Doctors
A. Classifying and Scoring Major Depressive Disorders
C. Medical Chatbot based on CBT
B. Parameter-Efficient Fine-Tuning Controllable Language Models

We aim to detect Depressive Severity with advanced techniques. (e.g., brain computer interface, AI)

We aim to fine-tune Controllable Language Models to generate Therapeutic Answers (Texts).

We aim to
(1) improve the performance of Pre-trained Large Language Models on downstream tasks by using Inhibition Adaption fine-tuning method that can inhibit redundant features.
(2) accordingly control the output of language models by NLP algorithms.

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-

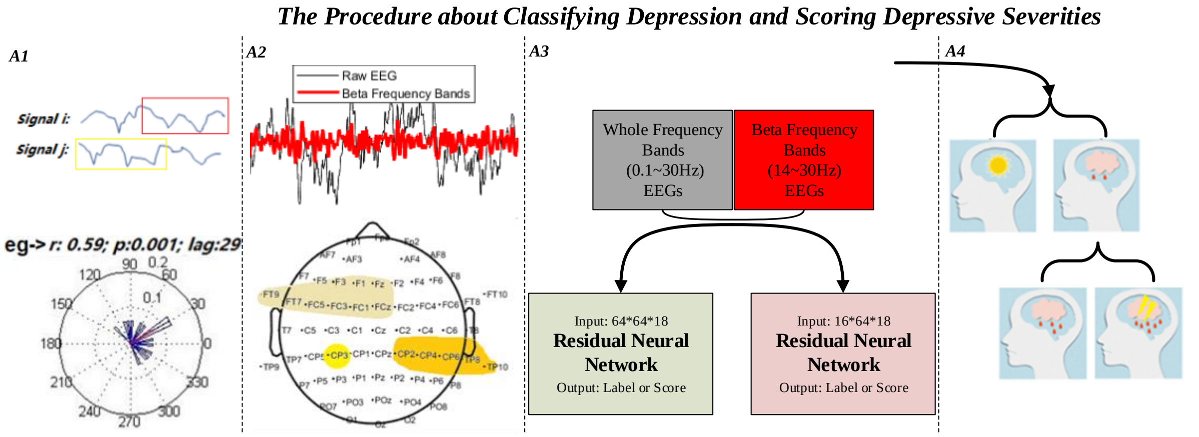

The Procedure about Classifying Depression and Scoring Depressive Severities
A1
Signal i:
Signal j:
eg-> r: 0.59; p:0.001; lag:29
A2
Raw EEG
Beta Frequency Bands
A3
Whole Frequency Bands (0.1~30Hz) EEGs
Beta Frequency Bands (14~30Hz) EEGs
Input: 64*64*18
Residual Neural Network
Output: Label or Score
Input: 16*64*18
Residual Neural Network
Output: Label or Score
A4

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. For example, the apparent age group estimation grades face images based on an ordinal scale: ”infants”, ”children”, ”teenagers”, ”youth”, ”young adults”, ”adults”, ”middle-aged”, and ”aged”.

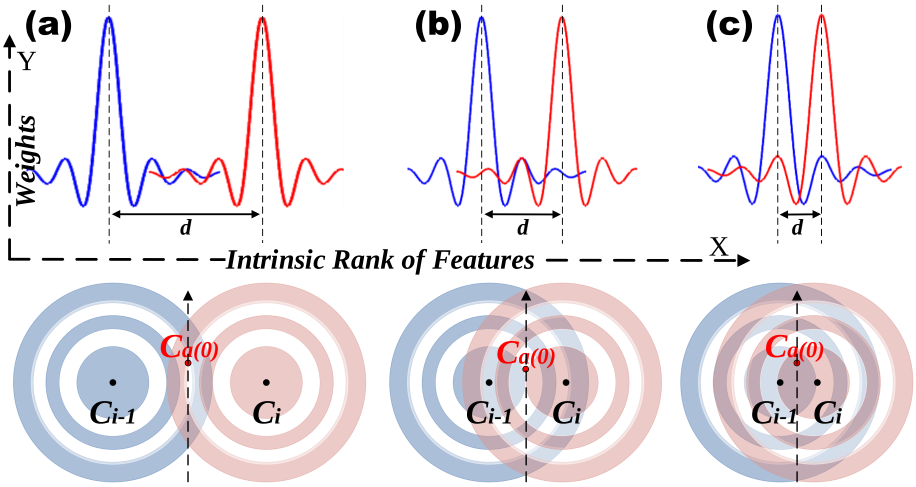


Figure 1.3: The challenge of ordinal image classification (or scoring). The X-axis denotes the intrinsic rank of features, and the Y-axis denotes the weights of models.

Ordinal classification can be viewed as a special case of metric regression, where the regression targets are discrete and finite. The differences in features between adjacent labels are not always equal to each other. The difference in facial features between ”infants” and ”children” being more obvious than facial features between ”young adults” and ”adults” 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. In Figure 1.3, $C_a(0)$ is the initial center of the overlapping feature. We assume that the features in ordinal images 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 medical ordinal images or 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 images, 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 image classification approach to analyzing ordinal data.

Ordinal image classification approaches or ordinal models [38] can be roughly divided into two aspects, Single Label Learning with Specific Loss (SLL-Loss) [39]–[42] and the

Label Distribution Based Learning (LDBL) [39], [43]–[50]. 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

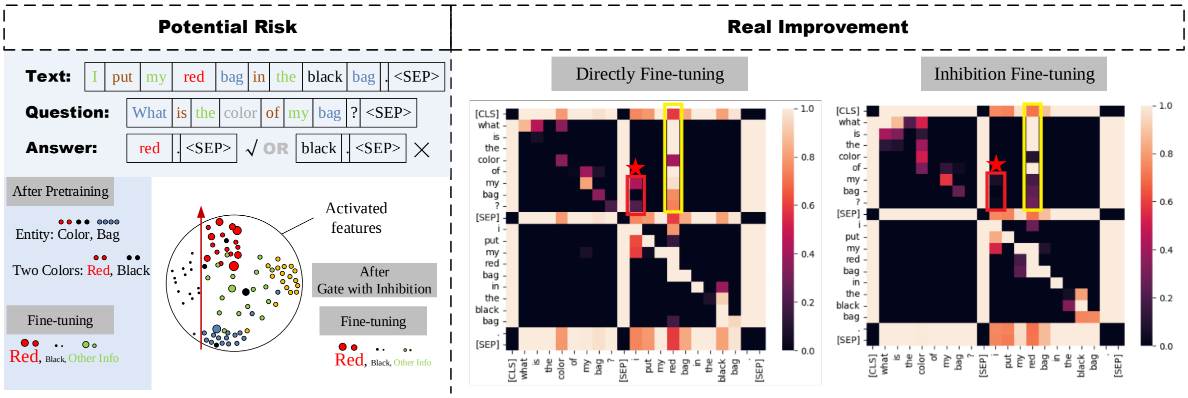


Figure 1.4: A practical example of InA is its use in the $BERT_{large}$ model, which has been fine-tuned under question-answering datasets.

Fine-tuning, the process of updating the parameters of pre-trained Language Model (LM)s, 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) ([51] (has successfully achieved this goal and addressed the inference latency problem, which helps extend model depth or reduce the usable sequence length of models ([52]–[54]) 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 ([51], [52], [54], [55]). 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 [51], 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 [56], chain-of-thoughts reasoning [57], and biomedical diagnosing [58]. Instruction-tuning of LLMs has enabled them to follow natural language instructions and perform real-world tasks [59]. 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 [60], and **(2)** supervised fine-tuning using public benchmarks and datasets augmented with manually or automatically generated instructions [61]. Reinforcement Learning on Human Feedback (RLHF) has proven to be an effective way to improve LLMs in various domains, such as medicine [62], knowledge graphs [63], and biomedical applications [64], but it comes with a high cost. Natural instructions

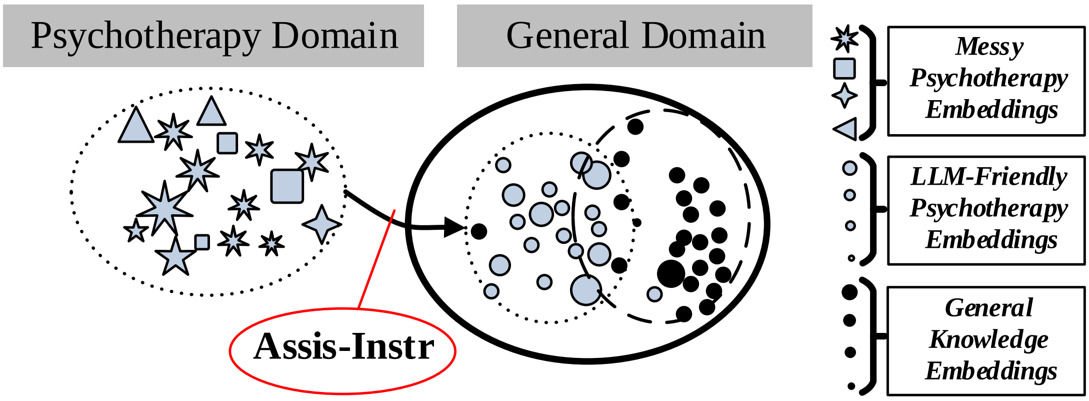

Psychotherapy Domain
General Domain
Assis-Instr
Messy Psychotherapy Embeddings
LLM-Friendly Psychotherapy Embeddings
General Knowledge Embeddings

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.

[59], and even un-natural instructions [65], can provide knowledge in multiple domains, but LLMs pre-trained on vast corpora (e.g., Llama1 [66] and Llama2 [67] containing books, common crawled conversations, arxiv articles, GitHub, C4, and Wikipedia data) still require additional professional knowledge, especially from domain experts. Self-Instruct tuning [68], [69] 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 the 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 tass which is also called an ordinal regression problem. Many published conventional methods treat age estimation as a traditional regression problem and make a strong assumption that each age dataset owns adequate instances to outline its distribution. Our FW-GPL aims to adaptively refine the age 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 age 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 facial age segments. It can produce equal performance when using the whole continuous facial age dataset. 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 on 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, 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 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].

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.

To overcome these three above challenges, we firstly assume that the delta and beta brain activities are connected to depression as previous studies [1], [7], [8], [70], [71] noted. Therefore, in an attempt to achieve early detection of depression, we have made an analysis of the delta and beta activities, as well as the corresponding brain networks which were provided as the visualization result. In our study, the Phase Synchrony Index (PSI) [1], [7], [8], [70], [71] was calculated to construct the brain functional networks. The electrodes and frequency bands were chosen based on different PSIs between depressive and healthy groups. Secondly, one classifier relying on ResNet [72] was designed to process selected EEG signals and detect depression. Additionally, we proposed one regression model relying on ResNet to score depressive severity. Both these two optimized ResNets on EEGs aims to accelerate the computing and diagnosing. This BCI system is expected to be used as a complementary tool to detect depression and monitor depressive severity, as well as a tool for evaluating conventional treatments in hospitals and clinics.

To detect depression from the healthy and, to score the depressive severity based on some specific psychological scales, we proposed this BCI system. The contributions of this paper are as follows: (**1**) We present the central-parietal increased delta deactivation accompanied by strong beta activation in the severe depression group under working memory tasks. (**2**) We propose one classification ResNet with the use of specific frequencies and brain regions which achieves better and more practical results in detecting the depressive from the healthy. (**3**) We also propose one regression ResNet with specific frequencies and brain regions for scoring the depressive severity based on two professional psychologists' score labels. The codes and corresponding documentation can be found here: `https://github.com/ChengKang520/Classifying-and-Scoring-MDD-BCI`.

2.2 Related Works

It is important to detect depression at the early stage, and not to delay proper treatments. Undetected depression can lead to long-term suffering and even suicide. Machine learning approaches can make early detection.

In recent years, medical images and methods relying on electrophysiological signals have been developed. Most of them have focused on extracting brain networks and using diagnostics models. Figure 1.2 illustrates the experimental flow. Our previous studies found that delta and beta brain activities are different from activities which are observed in the control group [1], [7], [8]. Based on these findings and inferences, a particularly designed system is illustrated in Figure 1.2. From **A1** to **A4**, **A1** presents the calculation of the PSI between two EEG signals. The formulas for calculating PSI can be denoted as:

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

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

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

where $\Delta\theta_{trialK}^{n\rightarrow m}$ 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 $r^{n\rightarrow m}$ is the mean value. We also denote it as PSI. Lastly, $lag^{n\rightarrow m}$ 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 [72] 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 relying on functional or structural brain networks are at the core of many mental health diagnosis methods, such as brain-connected networks for identifying bipolar disorders [73], [74] and schizophrenia [75], especially for detecting depression by EEGs [76]–[78]. To improve detection accuracy, researchers have focused on extracting useful information in the first pre-processing step.

In the first stage, when constructing functional brain networks or selecting obvious brain regions, some indexes were computed to estimate the connections or spectrums of brain regions. For example, during the resting state, one study used Adaboost classifiers to identify Cognitive Emotion Regulation Strategys (CERSs) with the help of spectral coherence [79]. Those spectrums of left and right frontal-prefrontal regions show obvious advantages to estimate the depression symptom under resting state [80]. The absolute power of the theta wave indicated a valid characteristic for discriminating depression. Relying on this finding, researchers used K-Nearest Neighbor (KNN) with 10-fold cross-validation to test the classification performance [76]. After the calculation of the relative wavelet energy and various entropy features by the Decomposed Discrete Wavelet Transform (DWT) coefficients on the EEG signals, a feed-forward ANNs was used to classify depression conditions [78]. A feature-level fusion approach was used to find powerful features, and then, traditional machine learning classifiers were utilized to detect depression by multimodal EEG data [77].

The brain networks have the attribute to show some well-known cognitive patterns, such as the abnormal cognitive control network of depressive patients [81], and they also can present the electrophysiological brain connections of some frequency bands (delta, theta, alpha and beta) [8]. According to brain oscillations in different frequencies, the PSI [1], [7], [8], [70], [71] was calculated to construct the brain functional networks. The PSI can reflect the degree of synchronization between two EEG channels, and after the calculation of correlation coefficients based on PSIs, an online clustering approach was used to construct convergent brain networks as described in previous studies [1], [7], [8], [70], [71]. Consequently, the Morlet's wavelet was utilized to calculate the time-frequency domain and the angle:

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

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

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 by machine learning approaches mainly contributes to the short time-consuming diagnosis. With the use of machine learning methods (Support Vector Machine (SVM), Adaptive Boosting (AdaBoost) and Random Forest (RF)), the most popular basic clinical techniques are Magnetic Resonance Imagings (MRIs) and EEGs. Some particularly selected channels were utilized before training tasks, as large amounts of irrelevant information will slow down training and make the model susceptible to overfitting [82]. The common flow of depression detection systems can be divided into the following three steps.

Step 1: Psychological paradigm. One study reported that adaptive dual n-back WM training can reduce subclinical anxiety and depression symptomology in adolescents [83]. During the learning processes of the WM capacity, WM was found to moderate the relation between the Brain Derived Neurotrophic Factor (BDNF) and psychotherapy outcome for depression [84]. These studies indicate that WM plays an important role in reflecting depression severity. Therefore, there are two reasons why we chose n-back tasks as the paradigm. **(1)** because most emotional tasks cannot control the affection intensity, we design this n-back paradigm and evaluate the WM capacity of participants by adjusting the n-back task (0-back is the base, and from 1 to 2,3,...,n, the difficulty of these n-back tasks will become higher); **(2)** to develop potential rehabilitation methods by using WM training for our future work.

Step 2: Feature extraction. Commonly, significant neuroimaging regions [81] and significant electrophysiological areas [19] should be extracted and fed into machine learning models, as well as some selected EEG channels contributed to depression classification during eye-close resting states or tasks completing states.

Step 3: Classifying and scoring depressive symptoms. In recent five years, researchers used traditional machine learning approaches with EEGs to identify depressed subjects, for examples, ANN[78], logistic regression[85], SVM[86], bagged tree[87] and

Convolution Neural Network (CNN)[19]. For classifying depression subjects, a deep CNN [19], a combined architecture of CNN and Long Short Term Memory (LSTM) [88], HybridEEGNet [89], and SNN [90] were used for discriminating depression under various cognitive or resting state tasks. Deep learning methods, especially CNN architectures, can automatically extract important features and can score depression severity within several psychological tasks. But for LSTM, one time-series model, which focuses on processing long time-series signals, it requires participants to complete long and huge continuous tasks.

For scoring the severity of depressive symptoms, one study used fMRI images and kernel partial least squares regression model, and the authors applied RMSE to evaluate the performance of their models [91]. Finally, we propose an approach based on beta EEG and particularly selected sixteen channels for classifying depression and scoring depressive symptoms. ResNets were selected because they can avoid gradient vanishing and go deeper with fewer parameters [72].

2.3 Methodology

2.3.1 Participants and EEGs Recording

The EEG signals were obtained from Shenzhen University, and Shenzhen Kangning Hospital, in Shenzhen, China. This study was approved by the ethics committee of Shenzhen Mental Health Center. This dataset consists of 52 healthy undergraduate dextromanual students ($6:4$ males to females; $Mean \pm SD = 20.4 \pm 9.7$ age distribution) and 48 depressed patients ($6:4$ males to females; $Mean \pm SD = 34.3 \pm 12.1$ age distribution). Strict selection and assessment procedures were employed as detailed elsewhere [1]. In both healthy and depressive groups, no medication was taken, and no personal or family history of psychiatric or neurological diseases was found before experiments. Based on using the SCID-CV [23], the 17-item HAMD, the depressed scores come from two professional clinical psychologists. The accepted depressive patients should be tested before the experiment by these two psychologies. If when two scores of one patient from these two psychologies could not be verified as the same level (one is mild, another is severe), this patient will be rejected to participate in this experiment. On the contrary, the average score will be computed to be the final score label. The depressive severity based on the HAMD and SCID-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 $none(non-depressed) = 0 \sim 7$ [92]. The difference between moderate depression and mild depression is small, and after pre-processing of EEG data, we found the data distribution is imbalance. To avoid these two potential risks that would affect the performance of our proposed models, we eventually reduced the number of categories to three. Based on these previous studies, in this system, three groups were selected: healthy controls ($non-depressed: 0 \sim 7$), depressed with low scores ($Score: 8 \sim 23$) and depressed with high scores ($Score :> 24$).

2.3.2 Working Memory Experiments

Following our previous studies [1], [8], this n-back experiment was developed under the E-Prime 5.0 environment. The letter variant version of the n-back tasks was used in the experiment. 0-back tasks were set as the baseline, 1-back 2-back tasks were set as the WM load. These volunteers observed and responded to the black letter stimuli on the screen with a white background, and at the same time, they should press two buttons including the index finger for matching stimulus and the middle finger for mismatching stimulus. Under the 0-back tasks, participants were required to identify a single pre-specified letter 'X' from the screen by pressing the matching button. Meanwhile, if they recognized a particular letter which matched the letter presented 2 trials before in the 2-back tasks, they should press the matching button. Presented letters were randomly selected from English consonants. This experiment was divided into three segments, and every segment has three tasks with a separately random arrangement of 0-back, 1-back and 2-back. We designed three tasks: 0-back, 1-back and 2-back. In each experiment segments, we randomly arranged the sequence of these three n-back tasks, because a predictable and fixed sequence design will affect the performance of participants when they implementing these WM tasks. Random arrangement of 0-back, 1-back and 2-back can avoid this potential risk. We set the duration of each given task as 75 seconds. All tasks should consist of a pseudo-random sequence of 30 consonants (10 targets and 20 nontargets). To avoid incorrect manipulation, as well as to provide enough time for reaction, letters were presented for 0.5 seconds and then disappeared in the following 2 seconds. Meanwhile, during every two parts, there are 45 seconds for participants to take a break. The behavioural performance was also recorded, for example, the reaction time and the response accuracy rate. Particularly, incorrect responses were excluded during the EEG analysis. After they convinced that there is no questions and every detail is clear, the warm-up tasks for guiding participants before the formal experiment would be ended.

2.3.3 Preprocessing of EEGs before Training

All procedures including EEG recording and preprocessing have been detailed in the previous studies [1]. Briefly, after (**1**) the removal of eye movements, (**2**) $0.16 \sim 30$

Hz (24dB/Octave) band-pass filtering, (**3**) artefact rejection and (**4**) baseline correction, the phase coherence calculation should be completed before the training tasks, because this study aims to develop an automatic system which can classify depression and score depressive severity by selected frequency bands and electrodes. Brain connection maps were constructed using the phase coherence method, as described in our previous studies [1], [7], [8], [70], [71]. The inputs are EEG signals from 64 or 16 channels during one type of task, and there are three different task types including 0-back, 1-back and 2-back.

2.3.4 Residual neural networks

In Figure 2.1, there are 64 channels that record EEGs, meanwhile, 2.5 seconds of EEG signals were collected. After that, a down-sampling process makes the data length from 2500 points to 1250 points. Thus, after discarding 98 points of the tail, we set the size of the input as $64 \times 64 \times 18$ in the first model. Then, two residual neural networks are used to train this EEG data (0-back, 1-back and 2-back tasks). Moreover, 16 selected electrodes based on the phase synchronization method were used in the second training phase. Therefore, we set the size of the input as $16 \times 64 \times 18$ for further training. The whole size of this EEG data is $22.5M$ 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 = $22.5M$). After we tested the CNNs with 6 residual blocks, the performance tends to be best, as the size of its parameters is 0.85M (properly selecting the size of the model's parameters can avoid overfitting, as well as insufficient fitting).

As the misdiagnosis rate of MDDs is widely recognized as 65.9% [14], we set the threshold of detecting rate as 70%. There are 60 trials that one participant should implement, and the depressive probability of one participant is the division product: (the number of trails whose predicted probability from the model is above 70%) ÷ (the total number of trails). Thus, if the predicted probability of one subject on one trial being depressive is above 70%, the system will classify him or her as 100% depressive on this trial. Finally, during one trial, if there are 33 subjects out of a total of 40 subjects whose probabilities of the ResNet classifier are equal to or greater than 70%, the accuracy rate of the model is 0.825 (33/40). Moreover, the second ResNet regression model outputs the score of depressive severity (as referring to the SCID-CV system and the HAMD score).

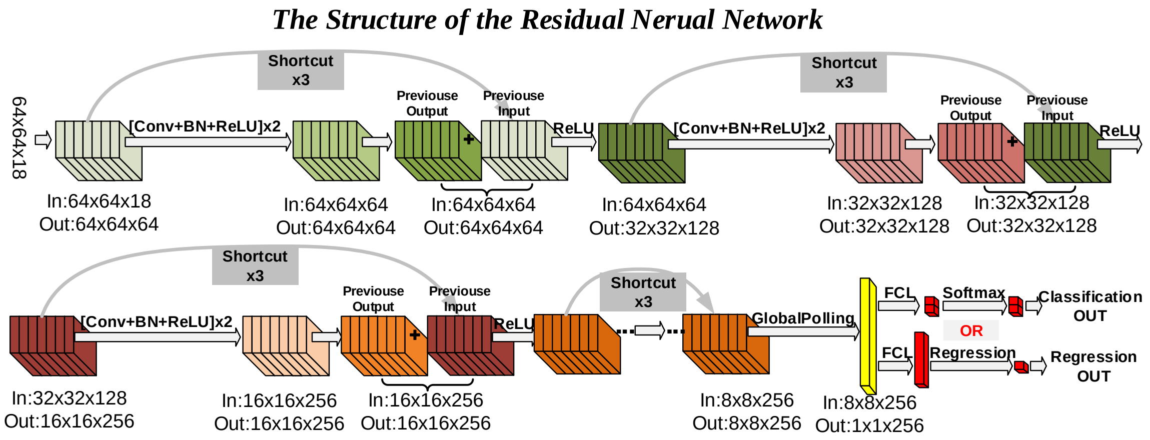


Figure 2.1: The structure of the constructed residual neural network. The input size is 64 × 64 × 18 or 16 × 64 × 18. Conv+BN+ReLu means the processing of convolution (Conv), batch normalization (BN) and rectified linear unit (ReLU). FCL is the fully connected layer. The shortcut is purely forward plus. ×3 means this block should be repeated triple times.

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 P = 0.061	Reaction Time P = 0.017	Accuracy Rate P <0.01	Reaction Time P <0.01	Accuracy Rate P <0.01	Reaction Time P = 0.053

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 ($P = 0.017$). 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 used the 0-back task as the "rest-state", and the 2-back task as the WM load. Thus, the PSI decrease refers to the weak neuronal activity in the corresponding regions and the inhibition trends returning to the ‘rest-state’. The PSI increase means the activation of the neuronal activity in the corresponding brain regions and strong WM related mechanisms.

Depending on two WM tasks (0-back and 2-back), the number of significantly connected pairs is presented in Figure 2.2. For the PSI decease, connections in the whole delta frequency components are the most significant different part among the three groups. The depressed group with high scores demonstrates the dominant whole theta frequency connections, but other frequency bands show no significant difference. When considering the PSI increase, the depressed group with high scores has the fewest delta, theta and alpha-connected pairs. Both two depressed groups demonstrate greater numbers of the whole beta connections, but the depressed group with low scores presents stronger connections in delta, theta and alpha bands when compared with the depressed group with high scores. After the comparison of the product (the number of significant pairs × the corresponding PSI values), which could represent significant PSI whole levels among three groups, t-value results deriving from the two-sample t-test are demonstrated in Figure 2.3 ($P < 0.01$). We marked the most significant frequency component in every histogram. Apart from Figure 2.3 B, which shows that the slightly obvious frequency part locates on delta bands ($P < 0.05$), the last three histograms show that the beta frequency activation represents the most significant difference.

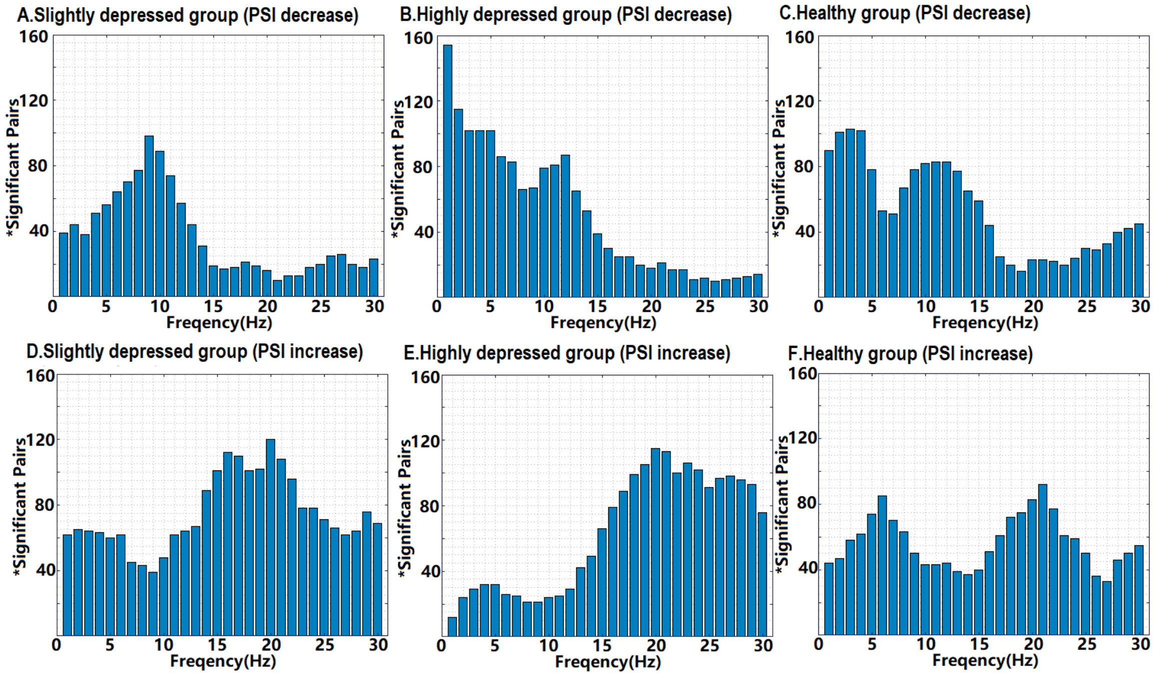


Figure 2.2: The number of the significant pairs in terms of the comparison between 2-back and 0-back tasks.

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-back	1-back	2-back	Best Result
(>0.70)	0.457±0.063	0.429±0.100	0.514±0.164	0.734 in 2-back
Score Difference	0-back	1-back	2-back	Best Result
(RMSE)	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-back	1-back	2-back	Best Result
(>0.70)	0.514±0.217	0.429±0.226	0.371±0.217	0.783 in 0-back
Score Difference	0-back	1-back	2-back	Best Result
(RMSE)	7.97±2.25	7.59±1.51	8.05±1.40	4.10 in 0-back

2.4.3 Clusters between these Three Groups

According to the PSI connections comparison, as shown in Figure 2.4A, the PSI decrease in the depressive group with lower scores contributes to few electrode connections, and it also presents the flat distribution in the beta frequency band in Figure 2.2. However, for the PSI increase in Figure 2.4B, the control group cannot generate one cluster, and the depressive group with low scores tends to gather the connected pairs in the left parietal and the left central regions (as shown in Cluster A). Accordingly, between the depressive group with high scores and the control group (the down panel in Figure 2.4C), in terms of the PSI decrease, the control group obtains fewer connected pairs mainly in the left frontal and the whole parietal areas (as shown in Cluster C), but the depressed group shows almost the whole cerebral connections apart from the occipital areas (as shown in Cluster B). Regarding the PSI increase, the depressed group with high scores presents the compact connecting pattern presenting the left frontal-central and right central-parietal regions, as well as the left frontal-temporal and the right temporal-parietal areas (as shown in Cluster D).

2.4.4 The result of classifying and scoring MDD patients

After these above pre-processing steps, there are no more than 60 trials for each subject, because we removed substandard trials. The result of the whole frequency band is shown in Table 2.2. Results of beta frequency bands are presented in Table 2.3. In the second

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-back	1-back	2-back	Best
(>0.70)	0.514±0.217	0.514±0.239	0.571±0.141	0.714 in 1-backs
Score Difference	0-back	1-back	2-back	Best
(RMSE)	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-back	1-back	2-back	Best
(>0.70)	0.452±0.302	0.409±0.222	0.414±0.367	0.833 (0-back)
Score Difference	0-back	1-back	2-back	Best
(RMSE)	8.12±3.38	8.07±3.44	7.74±3.66	3.02 (0-back)
Default ResNet (Size: 4.6M)				
Accuracy Rate	0-back	1-back	2-back	Best
(>0.70)	0.429±0.226	0.514±0.126	0.457±0.234	0.871 (2-back)
Score Difference	0-back	1-back	2-back	Best
(RMSE)	7.97±3.57	7.83±3.31	7.59±3.83	2.80 (2-back)

model, we extended the system for classifying depression and scoring depressive severity by extracting the beta frequency band and 16 significant electrodes. The online clustering step based on PSIs generates Cluster A and Cluster D, and the most frequently connected electrodes - Fz, F1, F3, FCz, FC1, FC3, FC5, FT7, FT9, T7, CP3, CP2, CP4, CP6, TP8 and TP10 – both in Cluster A and Cluster D can contribute to improve the performance of classifying depression and scoring depressive severity. To avoid random result, we used 10-fold computing method to choose the best result. For example, in Table 2.4, the classification accuracy rate can reach 0.714 when using the whole frequency bands. But as shown in Table 2.5 when relying on the beta frequency bands, the accuracy rate can even reach 0.871. Eventually, based on 2-back tasks in beta frequency bands with the contribution of particularly selected channels, in terms of 10-fold testing, 0.871 is the maximum value. Regarding scoring depressive severity, although the smallest RMSE result is 2.8 in 2-back in Table 2.5 when depending on beta frequency bands and particularly selected channels, the whole performance of scoring the depressive severity in Table 2.5 is weaker than that in Table 2.4.

2.5 Discussion

In this study, deactivation means that the rest-state takes the dominant role, and activation presents the processing of the working memory. We found that the low depressive group presents weaker delta deactivations but stronger beta activations, while the high depressive group shows more obviously deactivated delta connections and more activated beta connections. Moreover, there are beta right central parietal functional connections appearing in depression patients when the depressive severity becomes severe. In addition, the beta frequency bands contribute to classifying depressive patients from healthy controls, and particularly selected channels tend to easily distinguish depressive patients. Relying on beta frequency bands can increase the possibilities for scoring the depressive severity, and these selected channels also show obvious scoring advantages under the beta frequency band.

2.5.1 Possible inducing reason for getting depression

As the depressive symptoms becoming more severe, depression patients show more obvious delta deactivations and beta activations, but no evidence presents the obvious theta and alpha activities. This is a coincidence that subjects who were infected by Human Herpesvirus 6 (HHV-6) could not show a relationship with theta and alpha EEG oscillations [93]. Moreover, for patients suffering the HHV-6 infection, after medical treatment and 14 days of improvement, their theta/delta EEG oscillations will become slower[94], which means theta/delta activities become weaker [95]. Human Betaherpesvirus 6B (HHV-6B) infection can increase the potential risk of mental disorders [96], especially depression [97]. Then, we can conclude that there is a potential relationship between HHV-6 and depression. Our following research would focus on how serious HHV-6 can induce depression.

2.5.2 Topological analysis

The approach of topological networks provides the comparison of different cognitive patterns. On the one hand, the phase coherence analysis shows that the depressive group tends to weaken the low-frequency WM activation, especially in the delta and theta frequency bands. When depressive symptoms are becoming severe from moderately depressed to severely depressed, the above finding becomes more obvious. On the other hand, in Figure 2.3 C and D, the beta WM activation of highly depressed patients present a significant difference when compared to beta WM activities of the slightly depressed group. Thus, the depressive group would gain stronger beta activations than the healthy controls, and highly depressive patients show more risk of suffering from this imbalance. Moreover, 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. During the implementation of WM tasks, the depressive group reported reduced frontal-midline theta power and increased occipital upper alpha power during WM encoding [98], and this similar research perhaps provides evidence that depressive patients present abnormal brain activities in all frequency bands. The beta-frequency topological structure (Cluster D in Figure 2.4) of the highly depressive patients shows the extra central-parietal WM activation when compared with that (Cluster A in Figure 2.4) of the slightly depressive patients. This may correspond to the findings that only MDDs are characterized by unique EEG oscillations in beta frequencies. EEG beta oscillations are dominant in relation to delta, theta, and alpha when compared with healthy subjects [99], [100], and moreover, the high beta coherence is relative to connections within and between Dorsolateral Prefrontal Cortex (DLPFC) or temporal regions [80].

The increased delta deactivation during WM tasks represents the low WM loads, and this could be related to the resting recovery mechanism from cognitive maintenance. Considering the difference between Cluster B and Cluster C (Figure 2.4, panel C), as well as the pairs increase in the delta band (Figure 2.2), in terms of the condition when depressive symptoms are becoming serious, the same climbing trend of the delta deactivation was also seen in a neuromodulation therapy study [101]. Although connections of the highly depressive group in the PSI decrease show no obvious significance, the whole cerebral delta connections (Cluster B) indicate that they need more brain areas to implement WM deactivation than the control group (Cluster C). The research [101] also found that beta and gamma power increases at the Left-Dorsolateral Prefrontal Cortex (L-DLPFC) were correlated with an improvement in depressive symptoms. Increased attentional processes were proved to be connected to oscillations of the beta and gamma bands [102], and this might proof that Cluster A and Cluster D appearing in beta oscillations could adjust the attention processing of depressive subjects. When compared with Figure 2.2D, decreased pairs of the alpha activation in Figure 2.2E present the similar evidence that greater reductions of upper alpha and gamma power during WM maintenance were relative to high depressive severity [98].

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

When distinguishing depressive patients from healthy controls by a ResNet classifier in this study, the strategy depending on the single beta frequency band shows that the accuracy rate is higher than using the whole four frequency bands. Furthermore, for scoring the depressive severity in the depression group, this system presents a suitable approach to quantizing the grade of depressive severity. This finding may indirectly suggest that the beta frequency has the advantage to identify depression patients when implementing WM tasks [8]. Additionally, beta frequency cerebral activities can offer a tool to detect depression, but they cannot help to improve the performance of scoring depressive severity.

However, in the beta band, the scoring result presents wider variances. The whole frequency bands should be considered to score depressive severity only. The reason why the average accuracy rate is not high is that there are only two psychologists who diagnosed the patients and provided the results. This will cause the instability of data, especially when using the probably misdiagnosed subjects to test deep learning models.

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)[77]	MDD = 86, HC = 92	10-fold	KNN + EEGs (Fp1, Fpz, Fp2)	Highest at 86.98%
Xiaowei Zhang et al (2020)[103]	MDD = 81, HC = 89	10-fold	CNN + EEGs + demographic	Average at 75.29%
Xiaowei Li et al (2019)[104]	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)[91]	MDD = 58, HC = 65	leave-one-out	PLS + sMRI	9.56
Benson et al (2012)[106]	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

2.5.4 State of the art for classifying depressive patients

Table 2.6 shows the significant advantage of this proposed method, and the highest accuracy rate of detecting depression can reach 87.1%. However, the whole performance of scoring the depressive severity in 2.5 is weaker than that in 2.4 which uses the whole frequency bands. We infer that the reason is the quality of data and the robustness of the proposed model is not strong enough. In terms of the average accuracy rate, the limitation of this proposed method is that it cannot provide stable results, and this approach relies on the psychological paradigms (n-back) which only can represent the brain function of the 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) [105]. Table 2.7 shows that under the leave-one-out cross-validation, the minimum RMSE can reach 2.50 [106], 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 proposed a BCI system including two models based on the ResNet architecture (1) to detect depression and (2) to score the depressive severity by using 16 particularly selected channels and beta frequency EEG signals. The ResNet classifier is mainly for detecting depressive subjects from healthy controls, and the ResNet regression model is aiming to grade the depressive severity. Specifically, the coherence analysis provides the significant frequency bands, as well as the identified brain functional networks of depressive patients. We proved that the beta frequency can contribute to detecting depression and scoring depressive severity. Particularly selected EEG channels to show a significant advantage for classifying depression.

Future works will mainly focus on (1) the construction of further advanced ANNs, (2) the EEG data acquisition and selection of depressive patients, (2) more proper experiments design, and (4) the estimation of the antidepressant drug treatment. We will also aim to figure out whether there is a strong connection between inducing factors of depression and HHV-6.

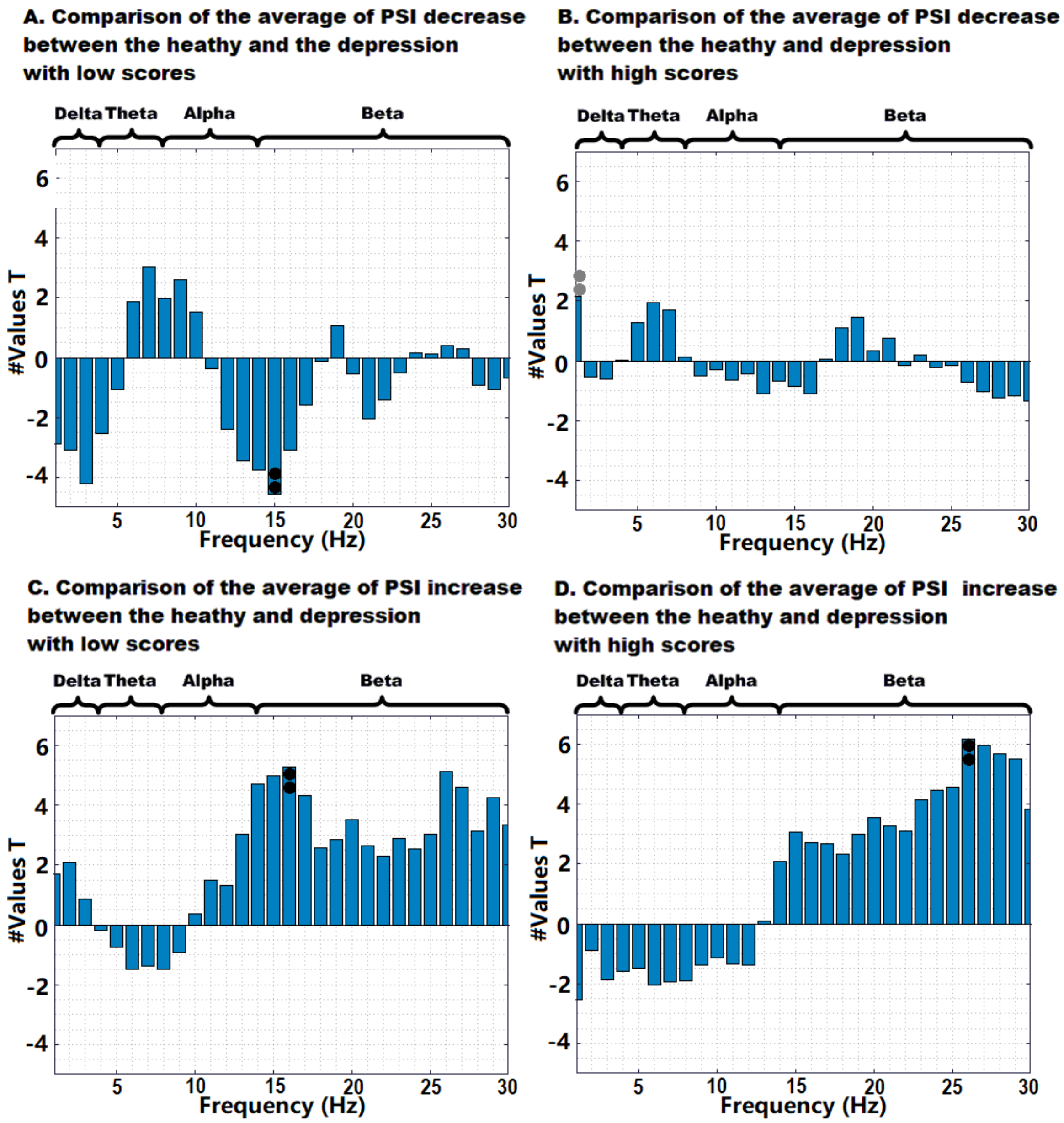


Figure 2.3: The t values (significant level) of the comparison between the depression group and the healthy control group.

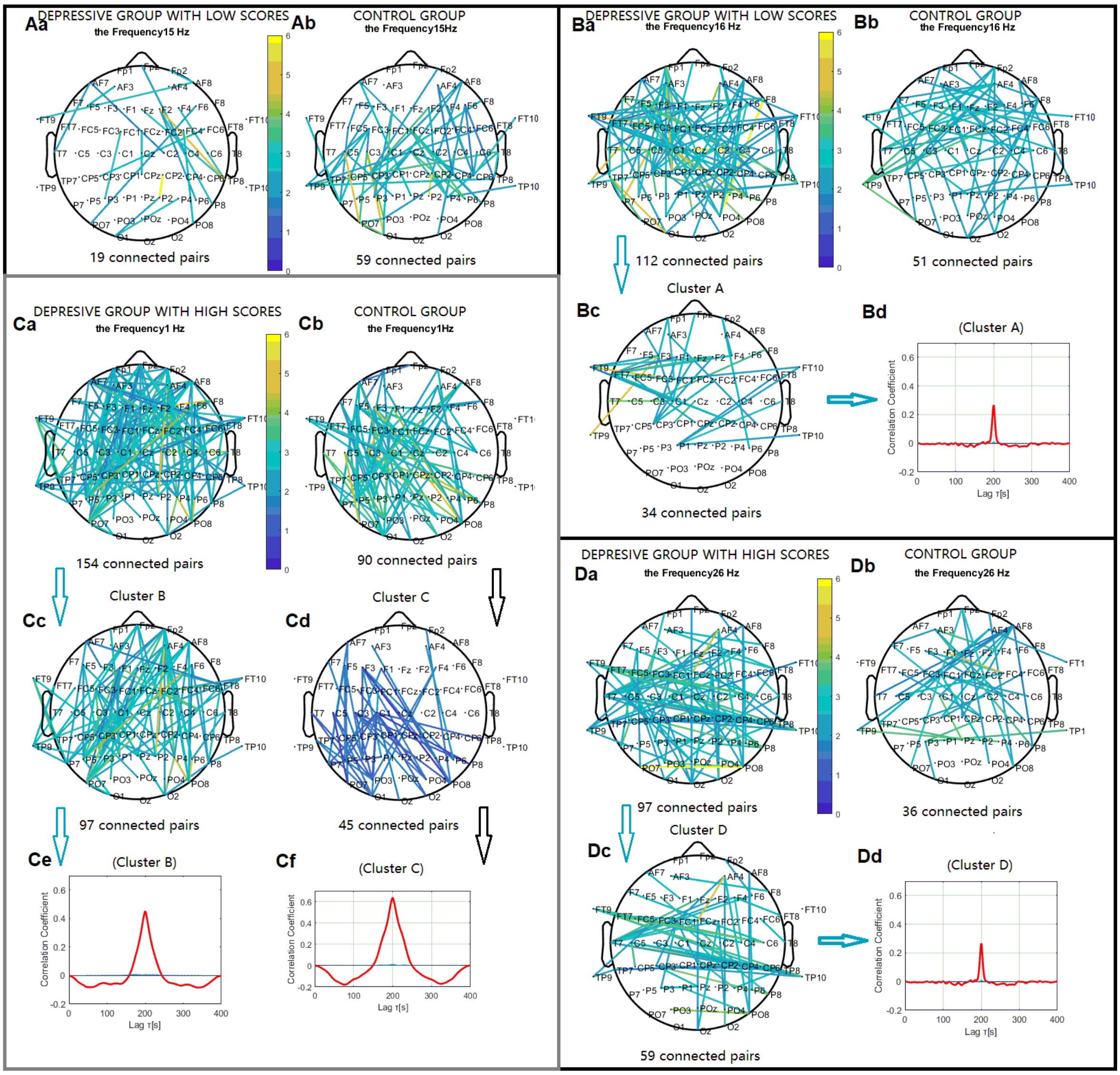


Figure 2.4: Clustering of some significantly increased and decreased phase synchronization indices mainly in beta bands for both the two depression groups and control groups. Lines in the up panel (panel A and B) respectively represent the significant PSI decrease and increase during the 2-back condition. 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 propose a FW-GPL for unconstrained facial age estimation, which is also called an ordinal regression problem. Many published conventional methods treat age estimation as a traditional regression problem and make a strong assumption that each age dataset owns adequate instances to outline its distribution. Our FW-GPL aims to adaptively refine the age 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 age 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 facial age segments. It can produce equal performance when using the whole continuous facial age dataset. Extensive experimental results on three facial ageing datasets and one ambiguous medical dataset demonstrate that our FW-GPL can achieve compelling performance compared to the SOTA.

3.1 Introduction

Ordinal regression, typically known as ordinal classification, is a supervised learning problem aiming to predict a discrete set of ordinal labels. Its main difference from the classification task is that the categories are related in a natural or implied order. For example, the apparent age group estimation grades face images based on an ordinal scale: "Infants", "Children", "Teenagers", "Youth", "Young adults", "Adults", "Middle-aged" and "Aged". Ordinal regression can be viewed as a special case of metric regression, where the regression targets are discrete and finite, and the differences in features between adjacent labels are not always equal to each other. There is an example which can explain it, saying the difference in facial features between "Infants" and "Children" is more obvious than of facial features between "Young adults" and "Adults". But if the ordinal relationship of labels is ignored, the ordinal regression problem will only become a simple multi-class classification issue. When learning with the long-tailed age data, a common problem is that the head classes usually dominate the training convergence. Therefore, the learned age classification model tends to perform better on head classes, whereas the performance degrades in tail classes. This quite motivates us to develop a robust facial age classification approach versus imbalanced ordinal data.

Facial age classification approaches could be roughly divided into two aspects, the SLL-Loss functions [39]–[42], [107] and the LDBL [39], [43]–[47], [107], [108]. SLL-Loss methods typically rely on independently processing a single facial image, and this ignores human face changing gradually with progressive ages, thus the facial appearance is usually indiscriminative or ambiguous at adjacent age classes. The LDBL methods tend to map the ordinal ground-truth learning based on a Gaussian or a Gaussian-Like label distribution. But the performance in such a long-tailed case where the conjugated features representation of ordinal neighbours or the rare feature of minority classes is suppressed by the majority classes.

To address the long-tailed and conjugated issue of the ordinal data, we propose the FW-GPL approach for the ordinal regression issue. It also aims to stretch semantic margins (or enlarge their inter-class variance) of neighbour classes which represent the shared features of neighbour classes. As shown in Figure 1.3, we assume that ordinal two neighbour classes have a closely shared feature region which can increase the difficulty of the ordinal classification task. A fuzzy window with Gaussian processed labels is carefully designed on the top of deep neural networks for limiting the effect of the semantic scoring trap and preserving the age-distribution information as well. Considering two technical aspects, our proposed FW-GPL is composed of two crucial branches including a difuzzifier window and a learning strategy of Gaussian processed labels. The difuzzifier window attempts to relieve and even eliminate the ambiguity of ordinal neighbours, and meanwhile, it tries to remain the internal ordinal age features which can represent the real facial age, and it also tries to narrow the classifier decision boundaries of tail classes by transferring the knowledge of head classes to tail classes.

Practically, the Gaussian processed labels allow the incorporation of prior knowledge (age Gaussian-Like distribution) to concentrate on the major age class and weaken the influence of remote neighbour classes. To validate the effectiveness of our proposed method, we perform extensive experiments on three widely-used face ageing datasets including Craniofacial Longitudinal Morphological Face Database (MORPH) II [109], Face and Gesture Recognition Research Network (FG-NET) [110] and Cross-Age Celebrity Dataset (CACD) [111], as well as one medical ordinal dataset: Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM). We achieve an abstract performance compared with the SOTA methods especially in processing the fragmentary samples because selecting the proper length of windows can shear off the influence of two tails. 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, $\{C_{i-a}, ..., C_{i-a+1}, C_{i-1}\}$ and $\{C_{i+1}, ..., C_{i+b-1}, C_{i+b}\}$, 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], [112], 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 [113]. There are some earlier studies working on constructing CNNs [114], [115], which have replaced the last layer of the ordinal classification model with a number of binary classifiers [116]. 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 [117]. 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.

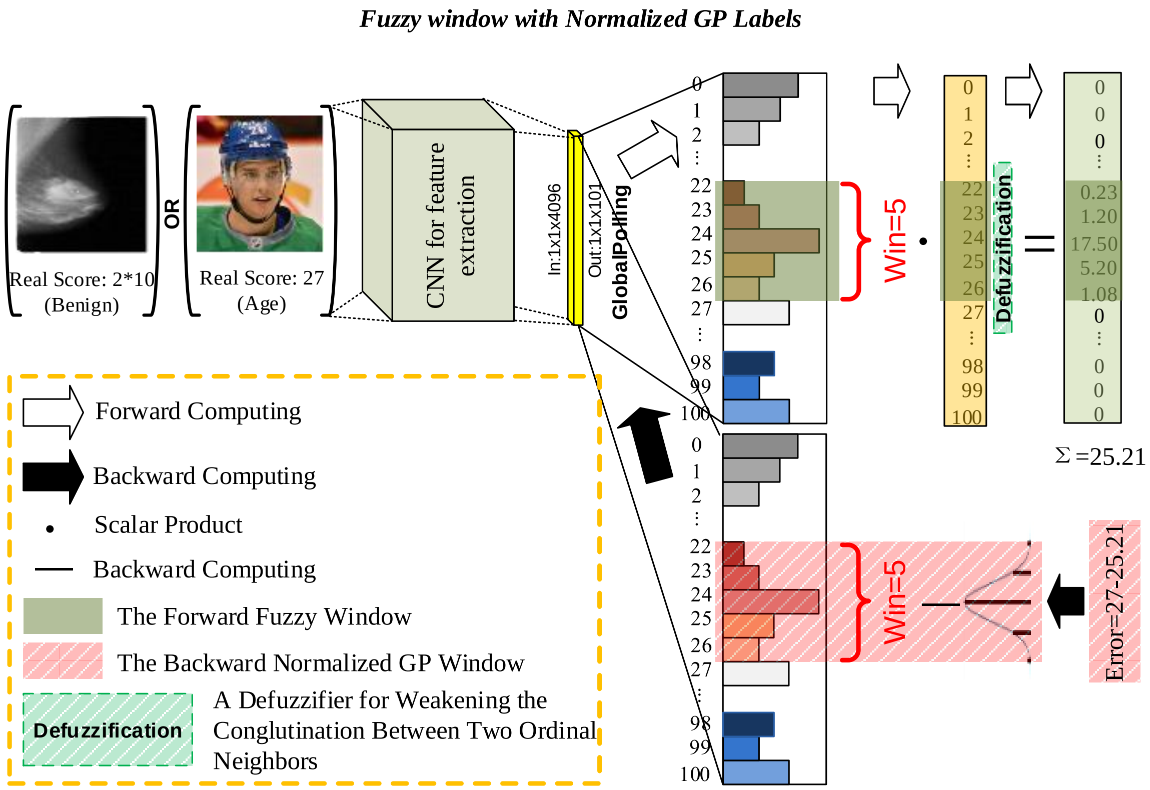


Figure 3.1: The proposed fuzzy window method (the length of the fuzzy window is 5) with the use of Gaussian processed labels for image scoring tasks.

3.2.2 Windows for Ordinal Classification

Moving Window Regression (MWR) [112] uses five neurons and a local window to estimate facial age. It proposes the notion of relative rank (ρ rank), a new order representation scheme for input and reference instances. This relative rank was estimated iteratively by selecting two reference instances to form a search window and then estimating the ρ rank within the window. In other words, MWR applies two overlapping windows with reference centers to limit the influence of the relative rank (or "intrinsic rank"), and it also uses a search process to find the most proper position of the centers to reduce the influence of the overlapping "rank". This has inspired us to develop a fuzzy window that can reduce 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 [118]. 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∼39 age group has a feature overlap with the 40∼59 age group), the training would become unstable.

When using fuzzy logic to solve the ordinal regression problem, one strategy achieved outstanding performance by extracting a set of fuzzy rules from an example set and using it as the basic model with a genetic algorithm [119]. Moreover, a method based on monotonicity indexes, an evolutionary fuzzy systems algorithm, was used for ordinal classification and ordinal regression tasks [120]. However, there is no common approach that can cover most ordinal image classification problems because most researchers prefer to develop particular methods or systems to target specific problems.

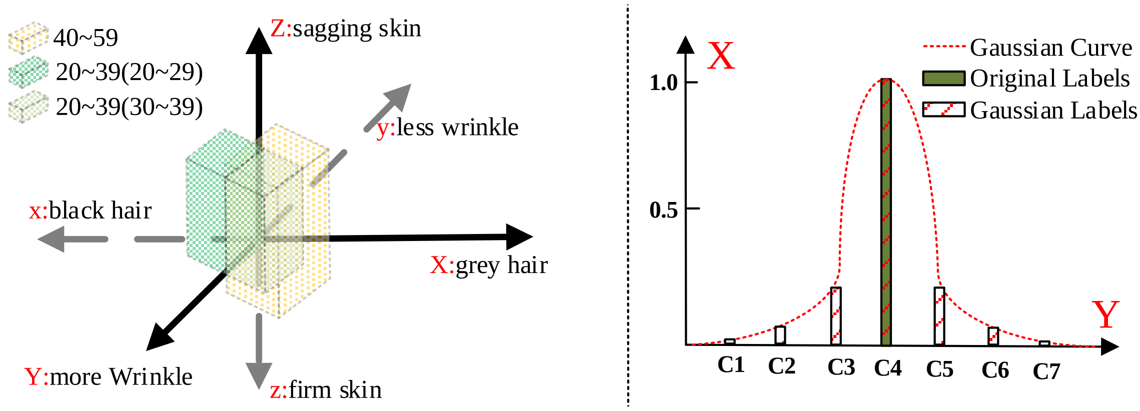


Figure 3.2: The left panel presents an example that shows overlapping features between two neighbor groups. The right panel shows the one-hot labels and the Gaussian processed labels.

Inspired by Deep Expectation (DEX), a fuzzy scoring method was used to reduce the influence of the tails and the shared features in each class and thus weaken the feature overlaps during the training steps [2]. Our previous work proposed a fuzzy window that focused on softly pulling the shared features to the optimal position by balancing the position of features with the distance away from the center. Under the conditions in Figure 1.3a, we set the length of the fuzzy window to 3, and the ascending (or descending) trend of the high score (or low score) was 1. Alternatively, under the condition of Figure 1.3c, in order to reduce the influence of the overlapping features, we set the length of the fuzzy window to 5. With this setting, the ascending or descending respective high or low score trend was 2. Eventually, the output value modified by the fuzzy window tended to slip forward to the global average position, and they optimized the redistributed probabilities as follows:

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

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. This type of label is traditionally used as a one-hot vector. For example, the encoding $h_i = [0; 1; 0]$ means that x_i is annotated to be the second class and $y_i = 2$. However, the undesirable application in classifying ordinal images is its disadvantage, as many of them are ambiguous or have an unclear "borderline." Thus, it brings difficulties to researchers as to which class they should belong to [121]. These hard labels tend to create an artificial gap that ponderously defines the borderline, and this instinctual drawback might then reduce the ability of the network to adapt [122].

Soft labels. By contrast, soft labels annotate categories by representing corresponding classes with a probability vector. For example, the encoding $h_i = [0.1; 0.7; 0.2]$ indicates that $P(Y = 2|X = x_i) = 0.7$. Thus, the value of the true label has been switched from 1 to 0.7 [121]. Instead of using a single bit, soft labels with probabilities can provide extra information to the training models [123]. Meanwhile, they have information inheritance, which can resist disturbance during inference [107], [124], [125].

Gaussian Processes. Gaussian process approaches for ordinal regression have been studied based on support vector machines [126], deep neural networks [44], and deep learning models with Gaussian distribution labels [127]–[129]. One partial label machine learning study used the Gaussian Processing (GP) approach to address and disambiguate the vague labeling information conveyed by the training data [45]. It assumed that there was already an unobservable latent function depending on the Gaussian process in the feature space of each class label. The essential problem, however, is that manually affected and annotated ambiguous labels have been ignored. The reason is that the Gaussian distribution, at times, could not really represent realistic labels without logic deblurring.

For facial-age detection, many technical articles, for example, Regression CNN (RCNN) [116], [130]–[132], Deep Label Distribution Learning (DLDL) [46], [107], and Deep Label Distribution Learning V2 (DLDL-V2) [39], reported that all of them trend towards the implicit use of the learning label distribution method, and data distribution is a Gaussian-like distribution.

The distribution of facial ages is represented by a Gaussian distribution for which a lookup table is generated beforehand to store multi-part integrals [133]. These integrals can explain the probability of whether an input image should belong to the true chronological age of a given person whose multiple age samples have been provided. In [47], a label distribution learning with a normal distribution variance σ was used and proposed $p_\mu(y, \sigma)$ 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, the most effective and popular method is using multiple binary classifiers to determine the ordinal category for each input (the K-rank approach) [39], [41], [107]. But the fundamental principle of this strategy is the consistency of the ordinal regression data [116]. In this section, we propose a simple and intuitive method that frames ordinal regression as a traditional classification problem, uses the Gaussian processed labels to stretch the shared features between two ordinal neighbors, and, finally, combines these Gaussian processed labels with a fuzzy window [2] to stabilize the weights on 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 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 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

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 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 $\frac{\partial E_m}{\partial W^t_{L-1}(\alpha,\beta))}$.

The difference between using GP labels and original labels is presented in Figure 3.3. We assume that there are two adjacent ordinal categories, $C_{i\text{-}1}$, and C_i, and, at the same time, shared quantized features, which are represented by the grey area. In the first round of Gradients Decent Direction (GDD)s, the original center of the shared quantized features is located at $C_a(0)$. If the true label is C_i, with the updating of the weights of models when using the original labels and back-propagation of errors, the initialized location of $C_a(0)$ would slide to $C_a(1)$ (see it in the Figure 3.3a). In the second GDDs round, if the true label is $C_{i\text{-}1}$ according to the vector direction of the pulling force, the center will slide to $C_a(2)$. Finally, after inducting the location of $C_a(2)$, the center of the shared features would stay close to either C_i or $C_{i\text{-}1}$ but not near the borderline.

Alternatively, when using the GP labels, the center of the shared features will fluctuate around the borderline of $\{C_i, C_{i\text{-}1}\}$. By using the original labels, the pulling force of back-propagation is unidirectional, which means the center of the shared features will move toward C_i or $C_{i\text{-}1}$ during every updating step. However, when using the GP labels, the pulling force of back-propagation is a resultant force generated from 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 [134]–[136]:

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

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

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

In order to make models more stable, practically, we should weaken the influence of gradients and use a lower learning rate or a smaller updating gradient. Additionally, when considering the optimal global strategy of only using a fully connected layer, the Fuzzy Fully Connected Layer (FFCL) can bring weaker influence into whole neural networks [2]. Thus, regardless of the strength of the pulling force (the gradient matrices in every layer), the center of the shared features can slide relatively smoothly to the optimal position. Then, this combined method is more beneficial to do classification in the output OR-CNN layer.

We use the DEX method as the base, and the true label y is quantized to different label

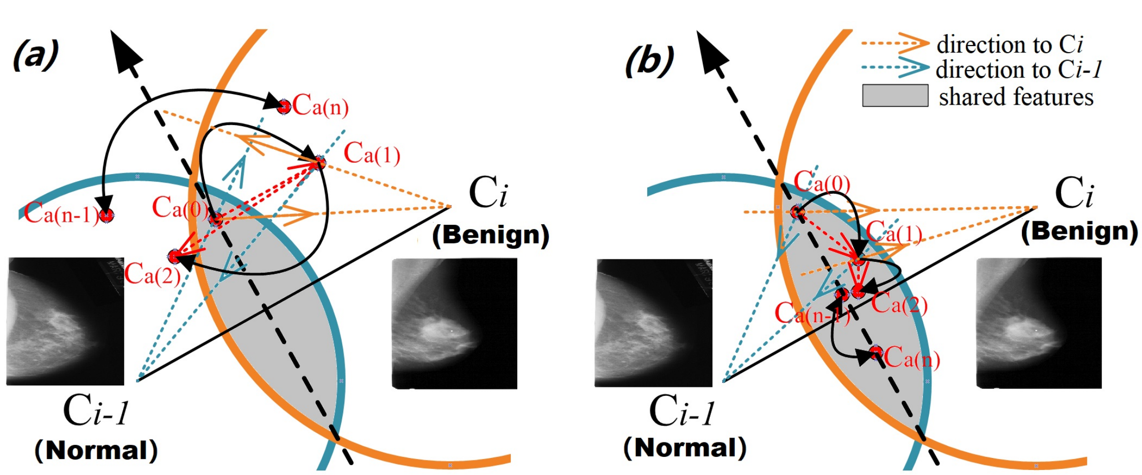


Figure 3.3: When two adjacent categories pull the center of the shared features, the resultant force decides where the center will finally stay. (**a**) When using one-hot labels, if the initial center of the shared features is $C_a(0)$, the resultant vector of the puling forces toward C_i and $C_{i\text{-}1}$ 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 $C_{i\text{-}1}$.

groups, which are treated as a class. To train DEX with fuzzy windows and normalized Gaussian processed labels, we replaced the expectation module—the last output layer—with fuzzy windows of different lengths, used a Gaussian function ($\sigma = 1/\sqrt{2\pi} \approx 0.4$) to process the ordinal labels, and, finally, modified the loss function with a typical cross-entropy loss. The back-propagation error between the output and the last layer after using the $Label_g$ was:

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

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

In Algorithm 1, we provide the pseudo-code of the fuzzy window with a normalized Gaussian processed label algorithm on processing the ordinal regression issue. The first step is to process the labels with Gaussian distribution. After setting the length of the Gaussian window, L_{Win}, the $Label_G$ can be calculated according to the Gaussian processing template in Table 3.1. However, if the Gaussian window reaches the head or the tail of the whole age sequence (0 or m), the element which is out-of-range (for example, 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 removed. The second step is to use fuzzy logic to eliminate the influence of overlapping features in ordinal neighbor classes. $\tilde{P}_i$ can be calculated out using Equation (3.3), and $\tilde{V}_o$ is able to be computed from Equation (3.4). During the inference time of the fuzzy window, an expected value, the sum of the multiplication of two elements—the position of the binary classifier and the prediction probability of this specific classifier—is used for the final estimation.

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 $Label_G$, $\tilde{V_o}$, $\tilde{P_i}$, $Error_G$;
```

3.4 Experiments

In this section, we introduce one medical image dataset Curated Breast Imaging Subset of Digital Database of Screening Mammography (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 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.

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)

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. CBIS-DDSM (Curated Breast Imaging Subset of DDSM) is a large collection of digitized film mammography images, which includes 3572 images referring to 2689 patient cases. According to BI-RADS, overall BI-RADS assessment from 0 to 5 has been described in this dataset, including BI-RADS score 0 (Incomplete cases), BI-RADS score 1 (Negative cases), BI-RADS score 2 (Benign cases), BI-RADS score 3 (Probably Benign cases), BI-RADS score 4 (Suspicious Abnormal cases) and BI-RADS score 5 (Highly Suspicious Malignant cases). This CBIS-DDSM dataset is available at `https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM`.

Depresion EEG. This EEG data has labels of depressive severity based on the Based on using the SCID-CV [23], the 17-item HAMD, the depressed scores come from two professional clinical psychologists. This dataset consists of 52 healthy undergraduate dextromanual students ($6:4$ males to females; $Mean \pm SD = 20.4 \pm 9.7$ age distribution) and 48 depressed patients ($6:4$ males to females; $Mean \pm SD = 34.3 \pm 12.1$ age distribution). Based on these previous studies [92], in this system, three groups were selected: healthy controls ($non-depressed : 0 \sim 7$), depressed with low scores ($Score : 8 \sim 23$) and depressed with high scores ($Score :> 24$). This dataset is available 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.

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 im-

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	⊗

ages of celebrities from IMDB (`www.imdb.com`) and Wikipedia (`https://en.wikipedia.org/`).

FG-NET. The Face and Gesture Recognition Research Network (FG-NET) [110] 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) [109] 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) [111] 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 [137], we computed the Mean Absolute Error (MAE) [138] and RMSE [139], 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 [115], Soft Stagewise Regression (SSR) [42], Mean Variance (MV) Loss [40], and Compact yet efficient Cascade Context-based Age Estimation (C3AE) [41], 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 $j = 1; j \leq epoch_N$ do
    if Current $Acc_j > Tre_{ratio} \times Acc_{average}$ 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.

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

Method	CNN + FCL	CNN + FFCL	CNN + W-GPL
Geras [140] (BI-RADS: 0/1/2)	68.8%	70.1%	70.3%
Akselrod-Ballin [141] (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%

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 [115]**	3.25	4.63	-	138M
	DEX * [115]	2.68	3.09	6.52	138M
	MV [40]	2.41	4.10	-	138M
	MV * [40]	2.16	2.68	-	138M
	DLDL-v2 [39]	1.969	-	-	138M
	FP-Age [142]	2.04	5.60	5.60	138M
	FP-Age * [142]	***1.90***	4.68	***4.33***	138M
	DRF [43]	2.80	3.47	5.63	-
	PML [143]	2.31	***2.16***	-	-
	JREAE [50]	2.71	3.390	4.596	-
	MWR [112]	*2.13*	-	5.68	-
	FW-GPL [Ours]	2.71	4.27	-	138M
	FW-GPL * [Ours]	2.24	2.73	6.10	138M
Compact	**ORCNN [40]**	3.27	6.44	-	479.7K
	MRCNN [40]	3.42	-	-	479.7K
	SSR [42]	3.16	-	-	40.9K
	C3AE [41]	2.78	4.09	-	39.7K
	C3AE * [41]	2.75	2.95	-	39.7K
	AVDL * [47]	2.37	2.32	-	11M
	MWR [112]	***2.00***	***2.23***	-	-
	FW-GPL[Ours]	2.72	3.71	-	40.9K

3.5.1 Scoring Breast Cancer Images

As we set the hyperparameter $L_{Win} = 3$ when scoring Breast Imaging-Reporting and Data System (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

With respect to scoring facial-age images, we set the number of neurons to 10 and the length of the window to 5. The results of a comparison between our model and SOTA models on three facial-age datasets are summarized in Table 3.5.

Compared with label distribution learning methods such as DLDL-V2 [39] and MV Loss [40], FW-GPL leverages a fixed pattern (Gaussian processed labels) to learn features that have considered the issue of age distribution. Our FW-GPL does not need to know beforehand the age distribution of the image data. Compared with some models that use special loss functions, we find that our FW-GPL, in particular, achieves competitive results compared with most SOTA methods, for example, MV Loss [40], SSR [42], and C3AE [41]. This is because the fuzzy window reduces the influence of the conjugation among neighboring ordinal categories, and it does not like DLDL-V2 [39], MV Loss [40], and SSR [42] considering the whole probability, or as with C3AE [41], focusing only on the two highest output probabilities. The second reason is that the Gaussian processed labels do not need to fit a proper hyperparameter σ [47] that should approximate the true age probability distribution. Compared with our FW-GPL, MWR [112] developed global and local relative ordinal regressors (ρ regressors) to predict ρ ranks within the entire and specific rank ranges. Furthermore, MWR first refined an initial search window, iteratively moved it by selecting two reference instances, and, lastly, estimated the ρ rank within the window.

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)[91]	MDD = 58, HC = 65	leave-one-out	PLS + sMRI	9.56	-
Benson et al (2012)[106]	MDD = 30, HC = 0	leave-one-out	RVR + MRI	2.50	-
		EEGs (Scenario)			
Hashempour et al (2022)[144]	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 [144] 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 [112]. 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**			
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**			
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 used two types of output layers (when $N = 100$ and $N = 10$) to analyze the performance of FW-GPL under different L_{Win}, with the results summarized in Tables 3.9 and 3.10. As the length of the half window L_{hWin} should be greater than the number of neurons in the output layer, we find that when the number of the neurons in the output layer is 100, a wider window ($L_{Win} = 50$ or $L_{Win} = 100$) gets better performance. Additionally, when the number of neurons in the output layer is 10, we can find the same trend. That is because the wider window can contain sufficient information to estimate the facial age. But if $N \leq L_{hWin}$, there is no improvement. These denote that a proper window length can improve the performance 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}.

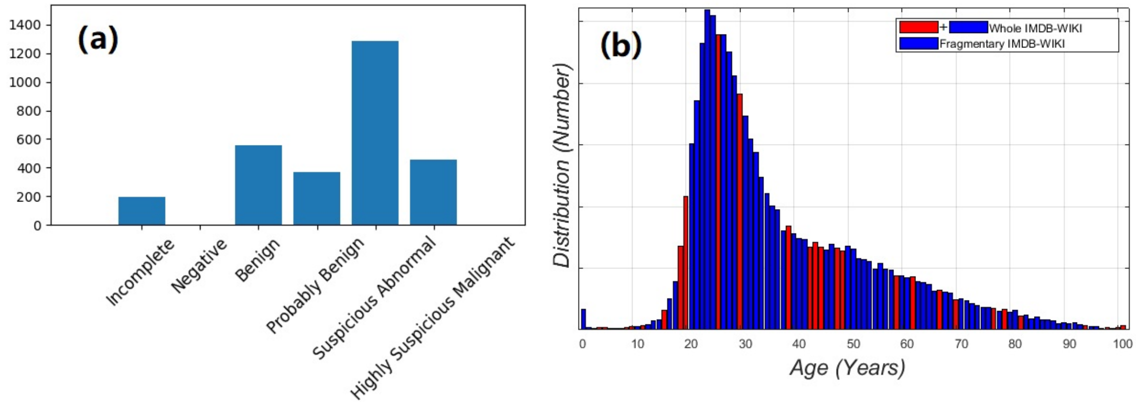


Figure 3.4: This figure shows the condition that the BI-RADS or the facial-age dataset is not consecutive. (**a**) The class distribution of CBIS-DDSM. (**b**) The age distribution of the IMDB-WIKI. The blue bars are the fragmentary IMDB-WIKI, whereas the red bars are manually removed.

Table 3.11: 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

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**

3.6.4 Advantage and Limitation

By directly facing the challenge of ordinal image classification, our method attempts to reduce the influence of the overlapping features. The length of the window controls the defuzzification of the ordinal neighbor categories. JREAE [50] used two covariance matrices to capture the underlying correlations from both aspects of input facial features and output age labels, but this family of methods (e.g., DRF [43] and AVDL [47]) should first take the age distribution of the dataset into account. This consideration is necessary because after fitting the distribution of the facial age dataset, there is an inevitable deviation between the real distribution of the age and the fitted one. To avoid such a problem, our method used a Gaussian distribution within the window to approximate the relationship between input facial features and output age labels. From Table 3.5, our method outperforms other LDBL methods and presents the advantage of using label distribution-based learning methods.

However, the disadvantage is that we only use a naive fuzzy logic window to leverage the challenge of ordinal image classification tasks. By adaptively adjusting the distance between the real age and the center of the moving window, MWR [112] moved the window to fit the ρ-ranks within entire and specific age ranks. Our method constrained the center of the window by using naive fuzzy logic to adjust the distribution of the facial age in the window. That would ignore the influence of the remote but highly related feature, which is beyond the window. Even though we tried to use longer windows, our method failed to overcome this problem.

3.7 Conclusions

In this paper, we have proposed a novel method for ordinal image scoring named fuzzy window with Gaussian processed label learning (FW-GPL). FW-GPL introduces a method to reduce the influence of the overlapping features between two ordinal neighbors. It achieves better performances than others on multiple age estimation datasets and one ambiguously annotated medical dataset. Our experiments also show that FW-GPL can process discontinuous ordinal regression by setting the proper length of the windows.

The idea of using fuzzy logic and a Gaussian process strategy to guide ordinal image classification is inspirational, and we will explore more possibilities for it. There are many directions for future work. (1) There are many other ordinal medical tasks, for example, scoring the severity of depression and grading the injury of spinal cords. We will use this method on such medical tasks in the coming research. (2) Our method cannot achieve the best SOTA result. We will try to overcome this challenge by infusing FW-GPL into other SOTA models. (3) To save the computing cost, we will fine-tune the pre-trained models which have been inserted with FW-GPL.

Chapter 4

Brain Networks of Maintenance, Inhibition and Disinhibition During Working Memory

WM – one of the brain ability thatmaintains information – can evaluate the function of brain. Activities related to memory sustention, inhibition and disinhibition have gathered significant attention for the basic neurocognitive architecture. Although researchers have proposed some brain models that attempt to explain the entire procedure of WM, little evidence can proof and describe it, and more particularly, regions and structures of maintenance, inhibition and disinhibition require more investigation. We used phase lock coherence and general partial directed coherence 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 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.

4.1 Introduction

WM has been defined as the capacity to guide behaviors [145], and it has also been associated with control of attention [146], and academic performance [147]. There is no consensus on neurocognitive architecture of WM, but the short-term maintenance of information is its basic conceptualization [148]. Many traditional WM paradigms with lower capacity were applied in many clinical populations to evaluate the performance of patients with mental impairments, including schizophrenia, stroke, traumatic brain injury and Attention Deficit-Hyperactivity Disorder (ADHD). There is an urgent and important clinical need (but unmet) for non-invasive measurements to evaluate WM activity and eventually to guide psychological intervention.

In this study, we actually (i) assessed the behavioral performance after subjects implementing n-back paradigms, (ii) examined brain networks of WM by phase-lock coherence and directional coherence after the 64 channels electroencephalogram (EEG) adaptively fitted and four sources generated to simulate cerebral internal communications, (iii) proposed our "neurocognitive architecture" of working memory based on region-to-region connections, and found the pathways of memory maintenance and lateral inhibition during WM. This present study illustrates the processes of WM and its relative regions through coherence in brain, and also presents a non-invasive assessment for functional networks during WM tasks in the healthy population.

A proposed neurocognitive architecture of working memory [148], already existing concept, consists of **1)** selective attention process, **2)** object information recognition and maintenance, **3)** rehearsal process, **4)** update and attention sustentation, and **5)** inhibition [148], [149]. This concept has offered an explanation for the procedure of WM by gathering other processing descriptions, and highlighted the concepts of memory maintenance and lateral inhibition [148]. These regions, including visual cortex, prefrontal cortex (PFC, mainly consists of posterior superior frontal gyrus and middle frontal gyrus), posterior parietal cortex (PPC, mainly seen in intraparietal and superior parietal cortex), and inferior temporal cortex have been brought together in visual WM paradigms [145], [150], [151]. This chapter illustrates the processes of WM and its relative regions through coherence in brain, and also presents a non-invasive assessment for functional networks during WM tasks in the healthy population. According to different neurocognitive stages, there are four major procedures during WM tasks. In this chapter, we actually:

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 suggested that PFC is critical for resilient information maintenance during WM tasks. Many meta-analyses have illustrated the left PFC (particularly ventral PFC) has been proved to be highly related with verbal WM tasks, whereas, the activation of right PFC (particularly dorsal PFC) is always followed by spatial WM tasks [152]–[155]. The association with PFC through lesion studies was confirmed that electrophysiological activities reveals neural connections in the PFC of monkeys [156], [157]. Moreover, fNIRS was used to evaluate the WM load by focusing on PFC blood activities [158]. Thus, these previous studies have established the importance of PFC in normal WM. Accompanied by the PFC, PPC was strongly recommended to be involved in WM tasks [159]. Spatial WM tasks, which tend to cover right hemisphere activation, activate bilateral parietal cortex [152], [154]. Then, a fMRI and Positron Emission Tomography (PET) study revealed that PFC could select the content representing in posterior regions [155]. However, several studies indicated that superior parietal cortex may be associated with executive function and selective attention control [160], [161]. Furthermore, previous research about the integrity of white matter pathways, illustrated the connections among PFC, parietal cortex, and temporal cortex during WM tasks [151], [162].

4.2.2 Pathway for Coding and Decoding

Working memory requires encoding and subsequent selection of appropriate content among distractors [163]. The interaction between PFC and PPC was proved to provide a top-down signal which can insulate stimulus-coding networks [164], [165]. Adaptive coding in PFC displayed the basic ability to classify learning tasks [166], [167], and one function, population coding, appearing in PFC neurons demonstrated the transition between several representational states relative to a delayed paired associates task [168]. The initial dynamic visual encoding in posterior brain regions and the encoding of selection rules in prefrontal cortex were presented by sources analysis, and these encoding and decoding components were particularly to maintain memory content [169]. Multivariate decoding and source analyses illustrated that selecting the memory content relies on prefrontal and parieto-occipital persistent oscillatory neural activity [169].

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 [148]. Particularly, fronto-parietal activity was examined to be relative to components of task-general processing, such as maintaining goals and task sets [170].

4.2.4 Pathway for Lateral Inhibition

During WM tasks, there are many items should be considered to be inhibited, as when the capacity of WM system is exceeded, inevitable inhibition can avoid the decays of persistent activity [145]. A winner-take-all dynamic model has been proposed to explain the strong lateral inhibition between memory representations, which indicates that the inhibition typically leaves only the winning representation active [148]. In order to integrate new and old information, two types of inhibition – 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) – were involved in WM tasks [171]. It was unclear how those differences regions were attributable to storage or executive components, and even communicating when implementing relative WM tasks [154]. To complete visuospatial WM tasks and tasks that require above components, these areas: superior parietal cortex, the posterior part of the superior frontal area and the middle frontal area, should be included [164], [165], [172]. However, although the relative active regions recently have been detected by above researchers, only few studies have been carrying on the architecture of inhibition and sustention among these areas, especially, the structure of loops to explain WM. In order to verify the "human neurocognitive architecture" of working memory, we used EEG sources and their connections to construct the communicational model based on these above cognitive components. In order to measure transmission of neural signals, various dynamic and statistical algorithms have been developed, more frequently, for example, directionality of neural information flow [173]–[175] and PLC relying on time lag [8], [176], [177]. Partial Directed Coherence (PDC) [178] and GPDC [179] have been proposed to analyze brain networks based on EEG studies. These two approaches (PLC and GPDC) have been generally proved to be significantly useful in structure systems [180], and in real organisms as well. For example, applicability/usefulness of EEG data were demonstrated on Parkinson's [175], Alzheimer's [181] disease patients, depression patients [7], [8], the hippocampal-prefrontal activation of monkeys [180], and kainic acid-anesthetized rats [182].

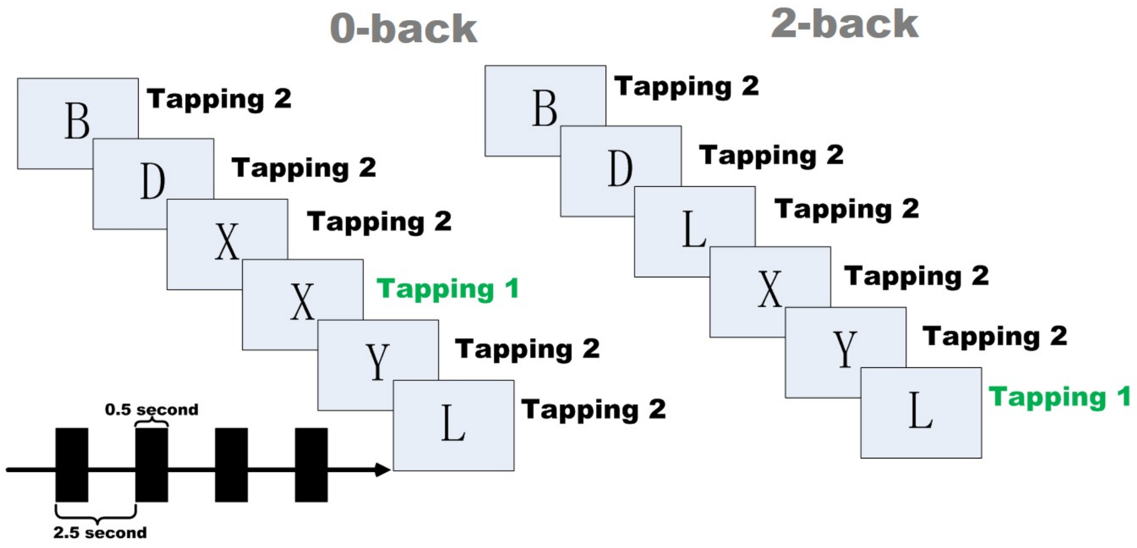


Figure 4.1: The experimental procedures with their timelines. Subjects should respond to stimuli by pressing the number key 1 with index finger for match (target stimulus) and pressing the number key 2 with middle finger for mismatch (nontarget stimulus).

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 used E-Prime 5.0 to design the n-back experiment in our study. The letter variant version of the n-back tasks was used in our experiment, including 0-back as a baseline and 2-back as the working memory load. Subjects should observe and respond to stimuli on screen by pressing buttons with index finger for match (target stimulus) and with middle finger for mismatch (nontarget stimulus). Subjects were asked to identify a single pre-specified letter ‘X’ during the 0-back tasks. Meanwhile, they also should recognize a particular letter which matched the letter presented 2 trials back in the 2-back tasks. Presented letters were randomly selected from English consonants (in Figure 4.1).

The entire experiment was divided into three parts, and each part included two 0-back tasks and two 2-back tasks, and the order of these tasks was randomly arranged. During either 0-back task or 2-back task, the whole duration of one given task was 75 seconds, and all tasks consisted of a pseudorandom sequence of 30 consonants (10 targets and 20 nontargets). To offer subjects time for reaction, letters were presented for 0.5 s, and then disappeared in the following 2 seconds. 45 seconds break was given between two parts. Every subject was required to answer as quickly and accurately as possible, and in order to analyze their behavioral performance, their reaction time and response accuracy were recorded, particularly, incorrect responses were excluded during the EEG analysis. Before the formal experiment, the subjects were guided to practice the tasks, therefore, they should repeat until they were convinced that every detail was clear.

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.

Data Preprocessing and Single-Trial Source Waveform Extraction

The averaged Evoked Related Potential (ERP) waveforms for single subject were calculated in 0-back and 2-back conditions respectively. The difference wave between the two conditions was calculated for single-subject and then by averaging all subject's difference in each corresponding channel, the entire representation of EEG was revealed to paint the performance of scalp topography (in Figure 4.2) before transferring it into source

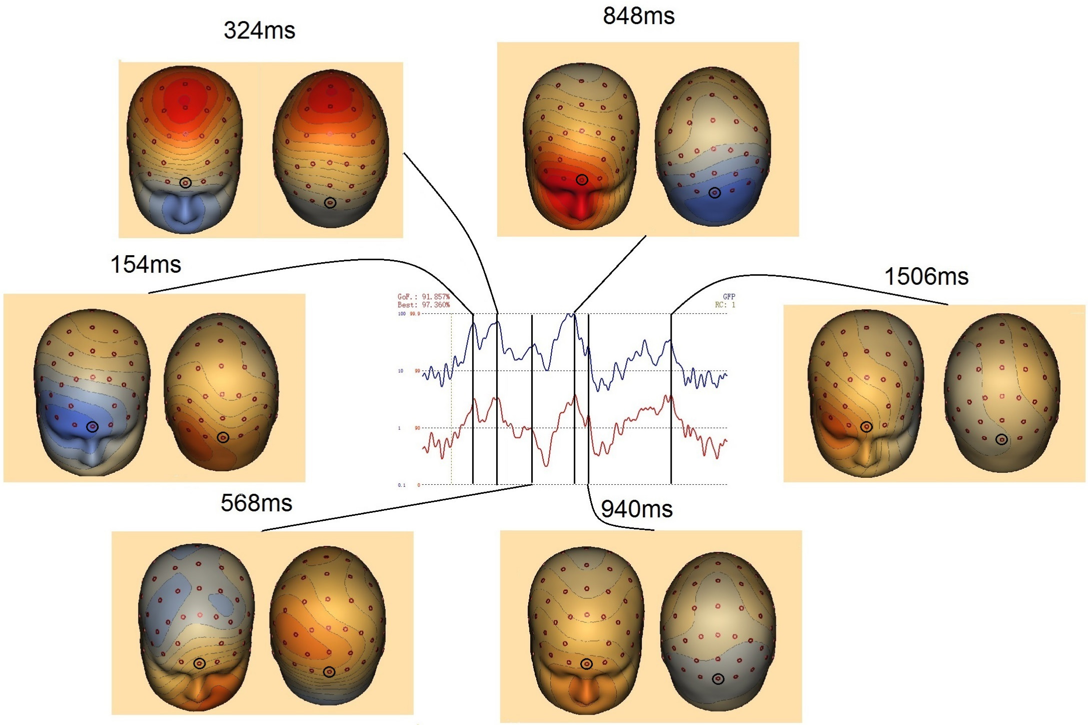


Figure 4.2: Illustration of brain altered scalp voltage maps of 2-back condition minus 0-back condition in the front and back hemispheres during different periods. The circled electrode sites are Fz and Oz. The GFP (the sum of squares of all channels, normalized to 100%) shown in the central is displayed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

waveforms (in Figure 4.3). Based on the fMRI findings of activations in the bilateral superior/inferior parietal lobules and bilateral inferior frontal gyri in the 2-back vs. 0-back contrast [183], a discrete model which has four sources was built to extract source waveforms. The reason why we constructed a Regional Source (RS) model instead of a dipole model is that a single RS composed of three mutually orthogonal dipoles can accurately present model activity of multiple gray matter patches with different orientation in a certain brain region [184], [185]. Meanwhile, the inter-hemispheric symmetry constraint for coordinates of the RS pair was used for source modeling. By feeding this type of difference waveforms between 2-back and 0-back conditions into Brain Electrical Source Analysis software (BESA 6.0), and according to the best correspondence between the recorded and estimated scalp distribution, we calculated the source configuration within a realistic head model and 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 (in Figure 4.3). Eventually, the resulting RS model was applied on EEG data in 0-back and 2-back tasks respectively, and it was used to extract their single-trial source waveform for each subject. Moreover, the obtained activity in the primary orientation of each RS (orientation 1 in Figure 4.3) was used for coherences analysis.

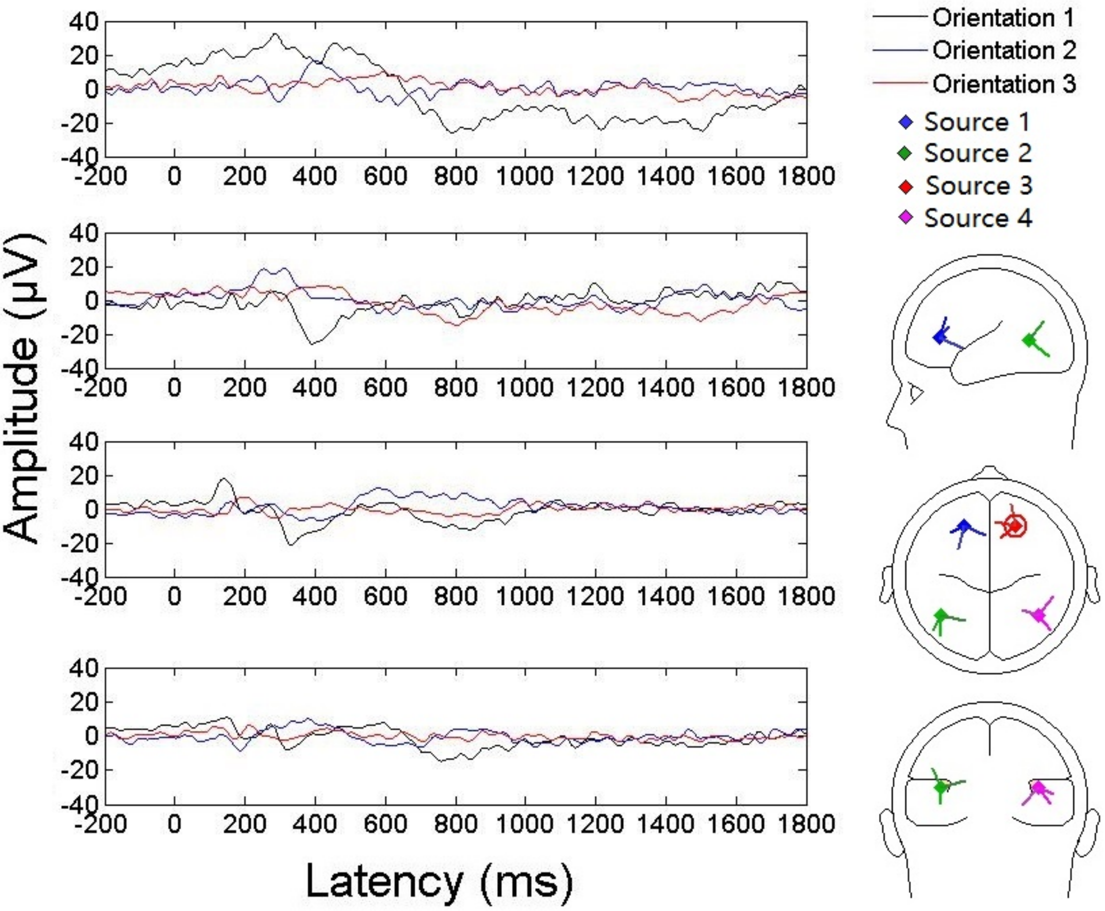


Figure 4.3: RSs and their corresponding time courses of the group average EEGs. The three directional time courses of RSs are displayed in the left panel, meanwhile, their locations and orientations of RSs are presented by using three head views in the right panel. Orientation 1 is the primary orientation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

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:

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

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], [180]:

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

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], [176], [177]. 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 [179] 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 [173]: **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 [173], we employed Kalman smoother method [186] 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 $(l \rightarrow m)_{th}$ 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:

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

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 resample 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 waveforms being group-averaged, we used 2-back to minus 0-back (baseline). We observed four obvious peaks presented in Figure 4.2. Early peaks appeared in 158 ms and 324 ms, and during these periods, the scalp topographic activity evolved from left temporo-occipital lobe to centroparietal lobe. According to the scalp topographic activity, prefrontal hyperactivity occurred during 844 ms and 1328 ms, which illustrates that activated areas were transferred into major frontal regions, meanwhile, the detailed information is that frontal potential reduced from 848 ms to 1328 ms, and the prefrontal, frontal and temporal lobes were activated.

4.4.3 Band-Specific Synchrony Reflects

We examined and verified the communication using phaselocking synchrony among these four sources in Figure 4.4. Before 700 ms (in Figure 4.4a, 4.4b), it revealed that the connection of S2-S3 shows a highly phase synchronized coherence with left PPC lagged behind right PFC (mean rel. $phase$ =-17.20, $p < 0.001$, $r = 0.943$, bootstrap test versus zero phase lag; in Figure 4.4a middle panel) mainly in late theta and early alpha bands ($6 \sim 11$ Hz). Meanwhile, for the posterior connection a strongly concentrated phase coherence (mean rel. $phase$ =-4.21, $p < 0.001$, $r = 0.875$, bootstrap test versus zero phase lag; in Figure 4.4a right panel) presented in late beta-band ($28 \sim 29$ Hz). For the phase-locked activities after 700 ms (in Figure 4.4c, 4.4d), there includes the front connection between S1 and S2 (mean rel. $phase$ =-17.91, $p < 0.001$, $r = 0.833$, bootstrap test versus zero phase lag; in 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; in Figure 4.4c left under panel) during 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;

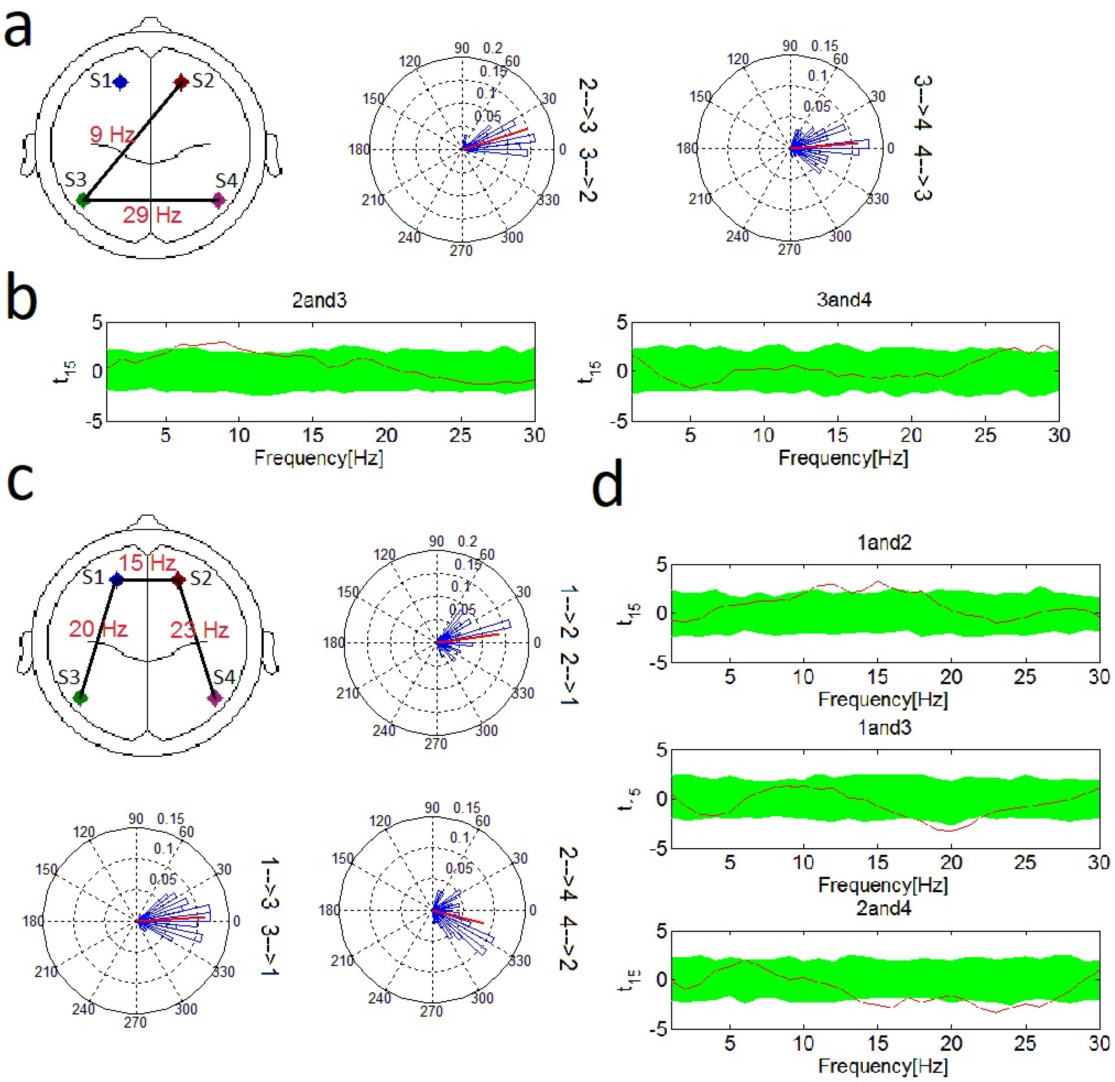


Figure 4.4: The phase locked connections among these four sources from 0 ms to 700 ms (**a**, **b**), and from 700 ms to 1600 ms (**c**, **d**). **(a)** Left panel shows the connections under some particular frequencies, and right panel presents circular statistic angles with their distribution, circular 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 ∼ 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.

in Figure 4.4c right under panel) during early and middle beta bands (14 ∼ 19 Hz, and 21 ∼ 26 Hz).

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

Referring to recent findings [145], [148] based on fMRI and electrophysiological research, some particular cognitive components were involved in Figure 4.6, for example, selective attention appearing in P300 duration [7], [187], verbal rehearsal [188], sustained activities [189], retrieve/readout [148], [190], pattern recognition [189], update and storage of memory [191], and lateral inhibition [145], [148]. Before responses, the posterior connection focused on the rouse of selective attention. Rehearsal and retrieve/readout in bilateral prefrontal regions, as well as sustained attention and pattern recognition between crossed right-prefrontal regions and left-parietal regions occurred during 500 ms ∼ 700 ms after approximately 250ms silent period. After responses, sustained attention and lateral inhibition in anteroposterior right hemisphere, accompanying with the updating and memory encoding in bilateral prefrontal regions, happened during 700 ms ∼ 900 ms. From 900 ms to 1100 ms, repetition of cognitive and memory components implemented to maintain brain activity in visual WM tasks. The last one took the charge of sustained attention to monitor targeted objects and lateral inhibition to avoid failure during 1100 ms and 1600 ms. Finally, we constructed and proposed a new neurocognitive architecture of WM processing in Figure 4.7, which can partially compensate the lacking explanation of WM. This architecture consists of several directed arrows and loops, and the indication of WM processes thereof.

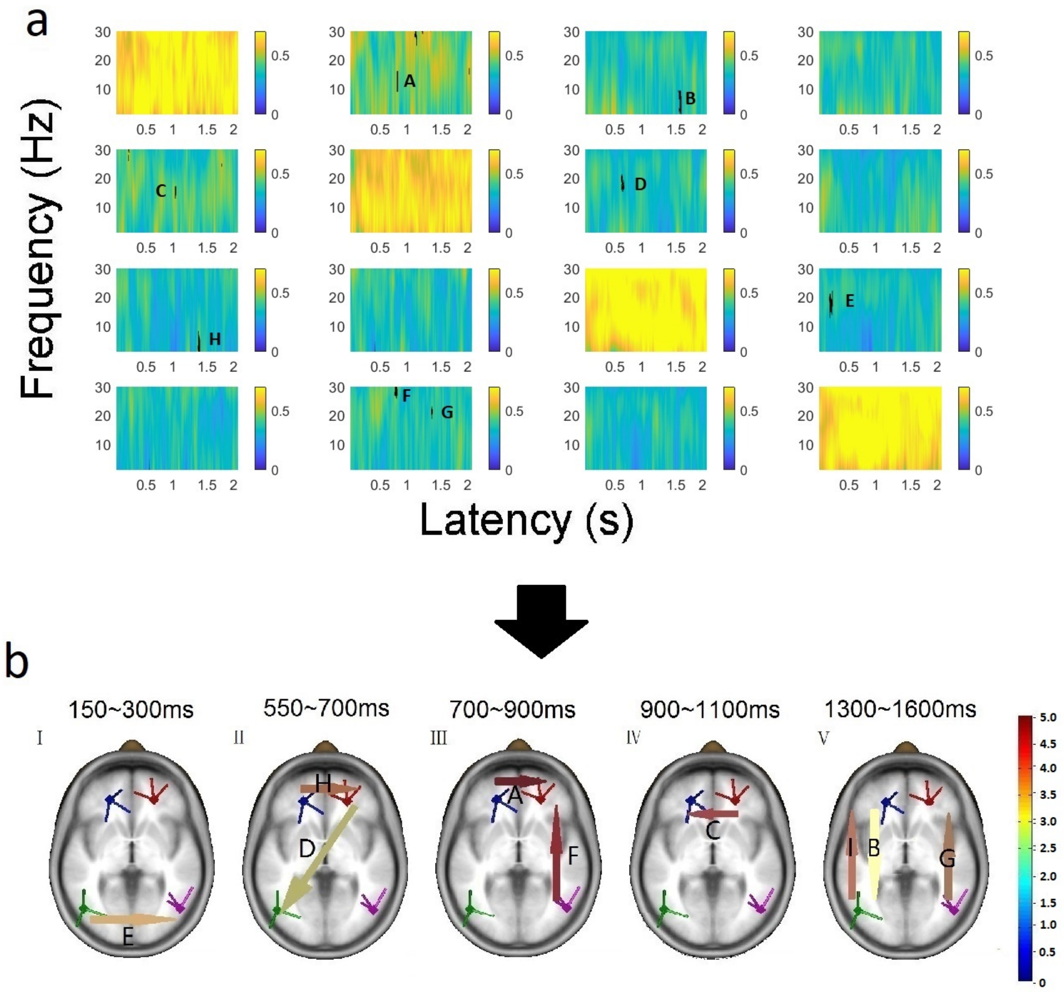


Figure 4.5: The directed connections based on the time-varying GPDC. **(a)** Time-frequency representations of the time-varying GPDC under 2-back tasks accompanying with significant grey blocks through two-sample t test between 0-back and 2-back. The bar presents the value of GPDC. **(b)** According 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.

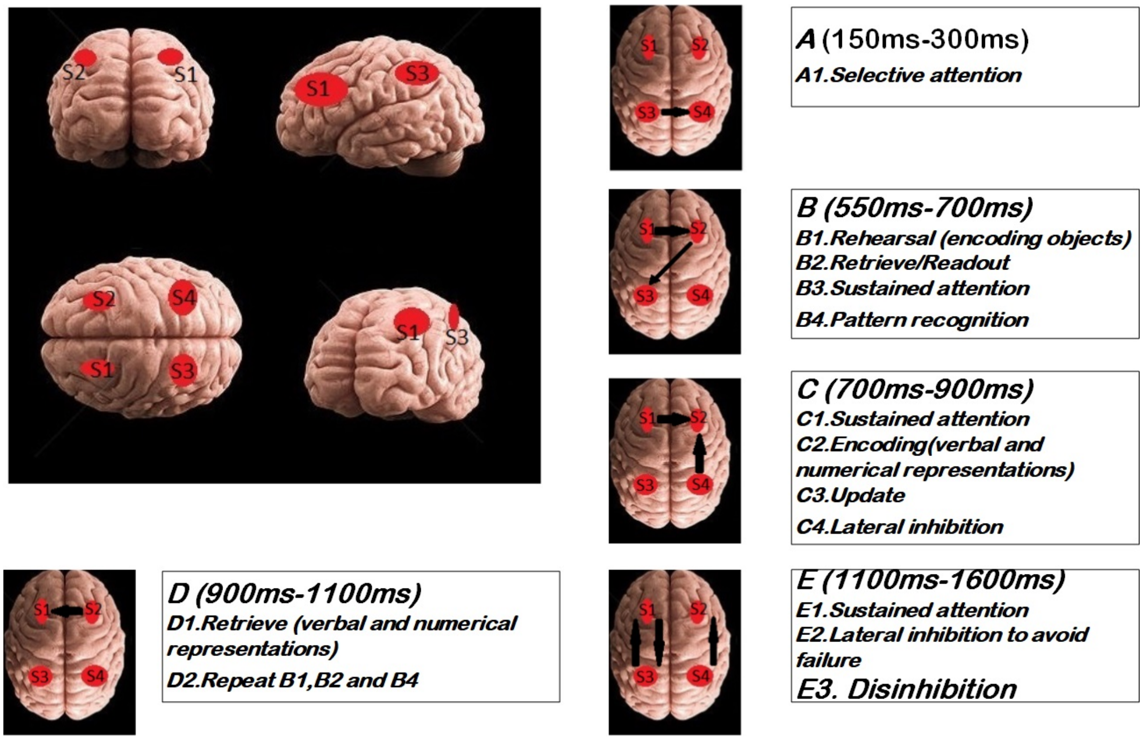


Figure 4.6: Schematic explanation of representations to brain networks during WM tasks. Left upper panel is the location illustration of four fitted sources. **A**~**E** present components relative to WM in terms of some specific neurocognitive processes. **A.** During this duration, selective attention 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 this traditional visual n-back paradigm, two coherence methods were used to construct the brain network under WM tasks. These adaptively fitted four sources mainly located on bilateral PFC and PPC areas, where functions relative to WM always appear in. Moreover, under the condition that 2-back minus 0-back, PLVs demonstrated the unoriented connections, while GPDCs illustrated the directional connections. Both of these two coherence methods presented the similar structure. Therefore, depending on above findings, we proposed a detailed model for WM processes, which consists of unrepeatable directional cognitive and executive connections, two regularly cognitive and memory maintenance cycles. Before responses, the targeted capital firstly triggered selective attention in parietal regions, and then it was encoded in visual cortex areas. The beta posterior connections in Figure 4.4a and wide beta directional causality in Figure

4.5b-I suggested the attention arouse, meanwhile, the target was locked on the screen. Compared with a speculation, beta oscillations in terms of selective attentional control may manage the attention and top-down procedure [192]. Eriksson suggested that a core circuit consisting of fronto-parietal cortical regions may sustain attention and implement rehearsal [148]. Fusing the main alpha coherence in Figure 4.4a and the beta directional connection in Figure 4.4b (D), it indicates the rehearsal simulation between right PFC and left PPC. Although these two methods presented different frequency bands when implementing rehearsal, this brain network might generate an early stage simulation derived from internal reasoning. Thus, the trigger activated the attention arouse, and then visual cortex encoded the target while transferring the representation information to the attention and rehearsal networks.

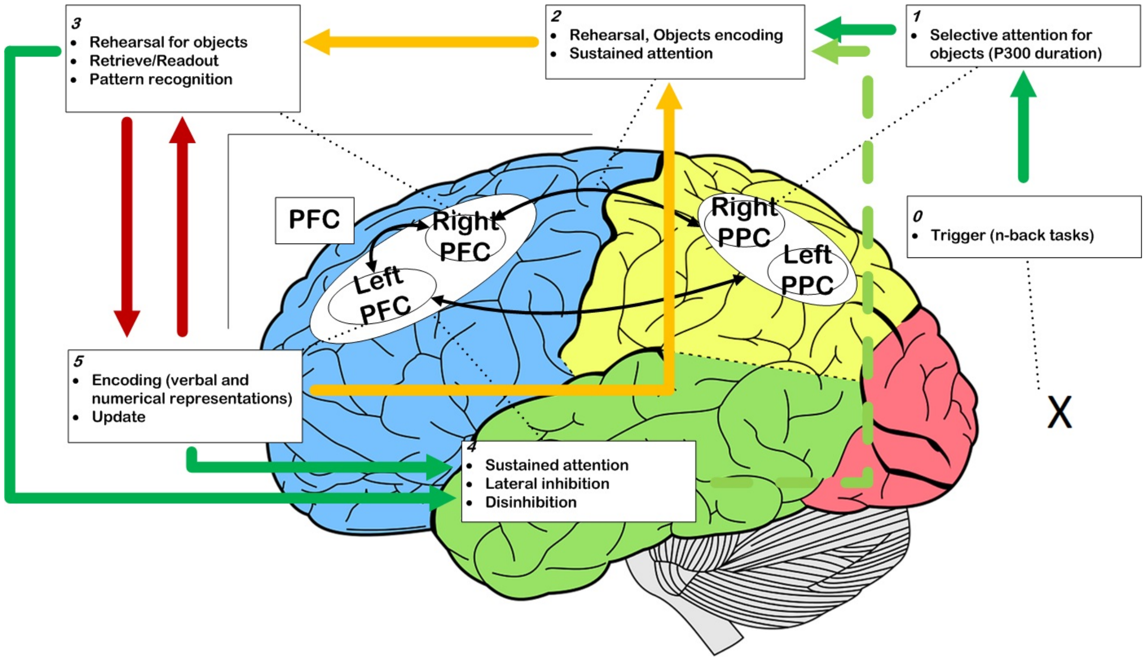


Figure 4.7: Summary of our proposed neurocognitive architecture for WM. **X** is the visual n-back trigger. Before responses, there are attention arouse link (**0-1-2**), activity maintenance loop (**2-3-2**). Attention arouse always accompanies with object encoding to recognize the type of objects. 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 reported to be relative to delay periods during which the activity profiles displayed across time [193]. Previous findings explored that frontal sustained brain responses during working memory tasks relate to selection mechanisms rather than the encoding of memory content [194]. Moreover, meta-analyses demonstrated that left PFC, especially ventral, is more associated with WM tasks, by contrary, the right PFC is more involved in spatial WM tasks [152]. During this type of n-back paradigm, directional connection D appeared in Figure 4.4b-II may explain why the information in right PFC will flow to left PPC. In our experiments, the right PFC was probably served as a buffer to store the buffer information and was read out to implement information selection and comparison during retrieval.

To keep the information or the activities in brain, shortterm maintenance needs sustention functions to support sustained brain activity during WM. From 300 ms to 550 ms, this approximately silent period has no significant activities, which partially indicates that our brain has a relatively stable activation when executing WM tasks. According to some cognitive findings [195], [196], we suppose that this type of "silent period" appeared from 300 ms to 550 ms seems more like to be a P300 component. Our experiment was not able to assess the mechanism of such a silent period during WM tasks, and we even cannot figure out the possibility that this relative silence to some extent may reflect the preparatory contribution for the following processes of high-level WM tasks, although some researchers believed that after a "silent period", followed by sustained top-down influences in brain, representations of information were transformed during WM to guide a final decision [169].

The red loop between bilateral PFCs in Figure 4.7 indicates the maintenance of memory information after responses. This red maintenance loop between left and right PFC may aim to activate the self-reminder function, and then either enhance the memory or guarantee necessary activity in brain. Recent fMRI analyses that requiring old and younger adults to implement WM tasks revealed that the older presented a weaker increase than the younger in BOLD signal in DLPFC when doing relative maintenance [197]. That study illustrated the requirement of both manipulation and maintenance, mainly because that subject pay attention to slip to the next position of the alphabet sequence and to maintain the result or the content of that maintenance operation in working memory. The yellow loop indicated the enhancement of short-term memory by repeatedly rewrite, comparison, and correction. Furthermore, the represented capital sequence during WM requires cognitive management to rank at least three capitals in brain. Therefore, although rehearsals can keep information in brain, the durable performance should be maintained through cyclic repetition. The most efficient way is possibly to rely on some auxiliary cortex regions, and the right PFC could act as a compensation area to schedule memory storing, to prepare temporary storage areas, and to ensure the sustained attention.

4.5.2 The Inhibition Loop During WM

WM was thought to be supported by the interaction between recurrent excitation among pyramidal neurons to sustain persistent activity by the delay (silent period), and lateral inhibition to mediate interneurons and to reduce the impact of external distraction [198]–[200]. Inhibition, especially the lateral inhibition, is the ability to reduce the activation of other functionally similar neurons within its local environment [201]. For filtering the distractor input, lateral inhibition would get hyper-polarized when background noise tended to influence neuronal firing [202]. Recent inferences highlighted the lateral inhibition within posterior areas, and the prefrontal cortex for resource limited descending [201]. Memory and positive brain activities maintenance essentially should control inessential information or functions to implement effective performance. Not only above activities require inhibition, but also the rouse and sustenance of attention. The lateral inhibition after the red and yellow loops are consistent with a recent theory that attention and working memory capacities are constrained by a flexible cortical connections to process overlapping inhibitory surrounds which can lead to internal competition for completing WM tasks [203].

Sometimes items can be forgotten due to the decreases of firing frequency by overall inhibition, which is not enough to maintain recurrent activity. Our experimental evidence weakly proved the prefrontal lateral inhibition during WM, particularly, after the delay period, it would intend to maintain memory representations and to avoid the switching of brain attention and task-focusing functions. However, we suppose that lateral inhibition more frequently causes deadlock cognitive stagnation for information maintenance, brain responses, and the rouse of following positive or negative neuron circuits.

To activate the following brain activities, disinhibition should be employed if the delay period increased, and this mechanism was evidenced by a Schizophrenia study during spatial WM experiment [202]. The effect of cortical disinhibition on WM may be the switch to activate the next loop of maintenance or brain activity, as the lack of disinhibition could generate disorder during the implement of one specific WM component.

4.5.3 Conclusion And Future Directions

We actually illustrated the phase lock and the directional connections among four adaptively fitted sources for noninvasive coherence analysis. By presenting the functional connections and relying on previous architecture of WM, we propose the detailed network of maintenance and inhibition, and evidenced the importance of disinhibition, especially indicated the partial functions of disinhibition in bilateral PPC regions. Modeling of this data reveals that **(i)** bilateral PFC and PPC play ahighly important role in WM tasks for attention, rehearsal, recognition, inhibition, and disinhibition; **(ii)** right PFC can provide an assist for left PFC to promote the high capacity implement of WM tasks; and **(iii)** after the inhibition for maintenance in brain (in one loop for enhance memory or keep positive activities), disinhibition would unlock the inhibition function to activate the following brain functions. In the future work, we will focus on the abnormal connections and unbalanced WM of 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) ([51](has successfully achieved this goal and addressed the inference latency problem, which helps extend model depth or reduce the usable sequence length of models ([52]–[54]) 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 ([51], [52], [54], [55]). 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 [204], and the concept of low 'intrinsic rank' in weight changes during model adaptation proposed by LoRA ([51]), 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 language models (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-

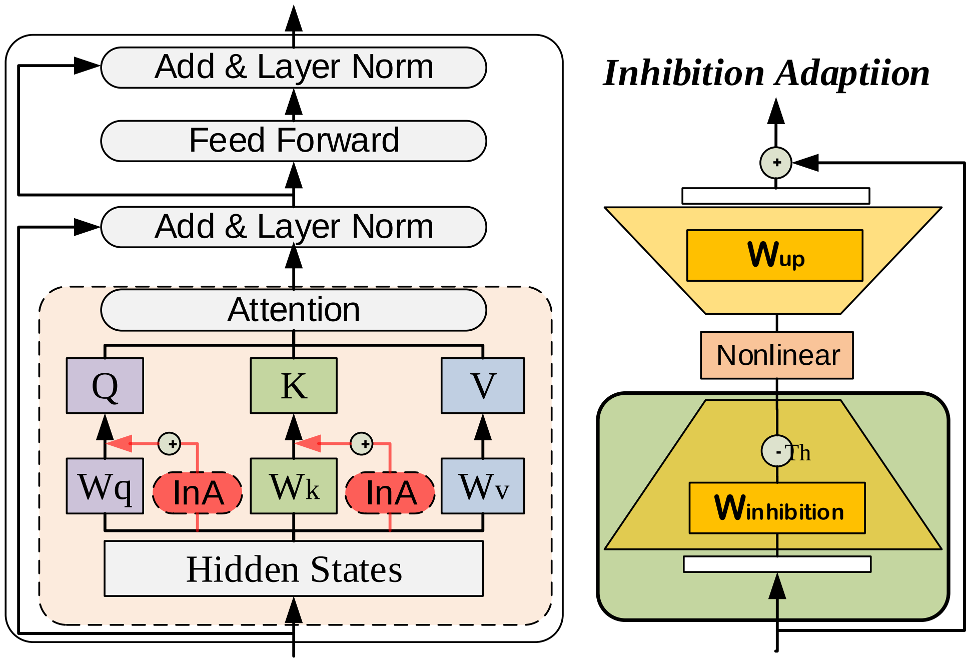


Figure 5.1: Illustration of the transformer architecture and our proposed parameter-efficient tuning method: inhibition adaption.

press task-irrelevant knowledge by subtracting a threshold.

5.2 Problem Statement

In the prior work of LoRA [51], 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 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 [205]. 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], [206], [207]). 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 ([206]). 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 Potential (EPSP)s.

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) ([208]), Parametrised ReLU ([209]), Exponential Linear Unit (ELU) ([210]), Self-Gated Activation Function (Switsh) ([211]), Gaussian Error Linear Unit (GeLU) ([212]), and Scaled Exponential Linear Unit (SELU) ([213]) 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 ([214]), 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

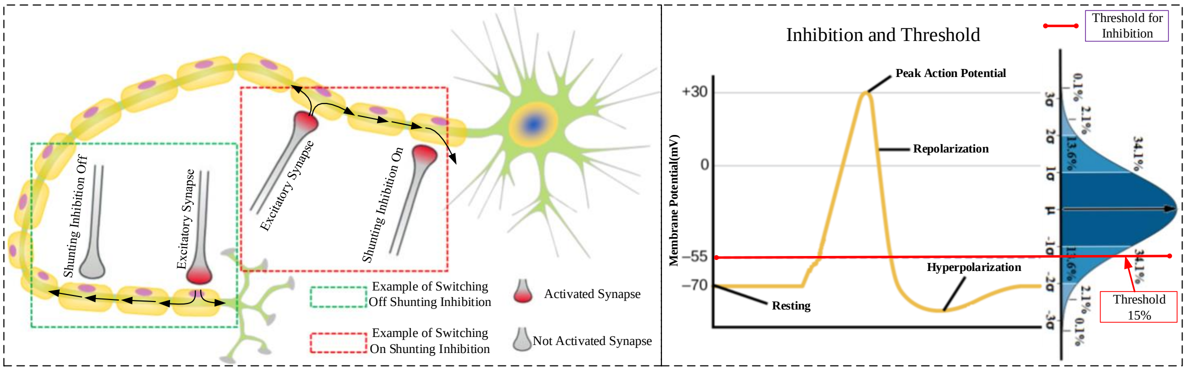


Figure 5.2: Inspiration from Neuroscience: Gate With Inhibition.

shape, context length and batch size), which was encouraged by the scaling law ([215]), has accelerated the capacity of various language models, such as BERT ([216]–[219]), RoBERTa ([220]), A Lite BERT (ALBERT) ([221]), DeBERTa ([222], [223]), sparse Switch-Transformer-1.6T ([224]) and Swin-Transformer ([225], [226]). 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 ([214], [219], [222], [225]):

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

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

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

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 ([216]). To maximise the downstream performance, variants of the vanilla Transformer (e.g., merely learning a subset of the parameters) compel practitioners to retrain allLM weights. For other conditional NLP tasks, such as question answering and dialogue generation, fine-tuning is also the prevalent paradigm ([222], [227]). 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 ([55]). 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 ([228]) 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 ([52]). 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 ([51]). 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 ([229]) which only fine-tunes bias vectors in the pre-trained model, diff-pruning ([230]) which learns a sparse parameter update vector, GLoRA ([231]) which generalizes the LoRA and QLoRA ([232]) 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 ([204], [233]). 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 ([234]). The firing thresholds are also fixed ([235]) or selected based on some heuristics ([233], [236]). The threshold was selected as the maximum preactivation of each layer in [233]. [236] 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 ([237]). 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 ([238]). Following prior studies of language models ([51], [230]), 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 $BERT-large$ ([221]), $RoBERTa-large$ ([220]) and $DeBERTa-large$ ([222], [223]). Whether natural language understanding, question answering or generation, specifically, the benchmark GLUE ([239]), SQuAD v1.1 ([240]), SQuAD v2.0 ([240]) and SWAG ([241]), we followed the adapter fine-tuning setup ([228]) 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 ([216]), RoBERTa ([220]) and DeBERTa ([222]), we adopt dynamic data batching. We also include span masking ([242]) as an additional masking strategy with a span size of up to three. For fine-tuning, we use Adam ([243]) 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 ([239]) in Table 5.3, as well as question answering – SQuAD v1.1 ([240]), SQuAD v2.0 ([244]) and Text Adversarial Generation: SWAG ([241]) 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 [55] 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 [216]	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 [220]	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]†[228]	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]†[55]	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]†[51]	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 [222]	304.0M	70.5	92.3	91.1	96.8	92.8	95.2	88.3	91.9	90.00
DeBERTaV3 [223]	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 ([216]) and DeBERTa ([223]) fine-tuning. Our settings for InA fine-tuning on $RoBERTa - large$ are, respectively, similar to the adaption fine-tuning method ([51], [228]).

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 [216]	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 [220]	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 [222]	304.0M	**95.5/90.1**	90.7/88.0	90.8
DeBERTaV3 [223]	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

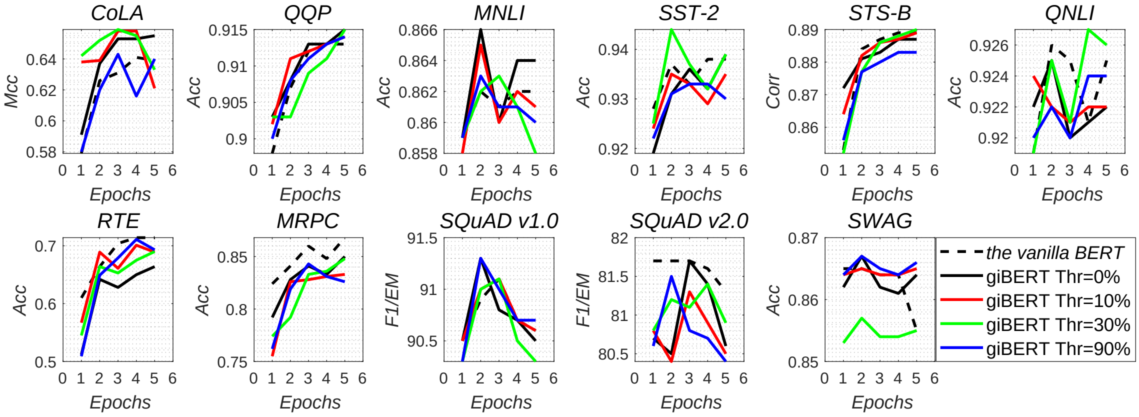


Figure 5.3: Plots of corresponding metrics according to the number of epochs on the validation split of GLUE, SQuAD v1.1, SQuAD v2.0 and SWAG. The giBERT means inserting InA (gate inhibition mechanism) into BERT.

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 ([240]), 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.

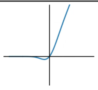
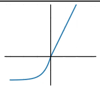
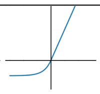
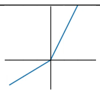
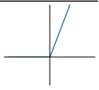
InA on the Multi-Choice Generation Task.

In Table 5.4, for the SWAG text generation dataset ([241]), 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	**95.8/89.3**	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	**95.8/89.5**	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

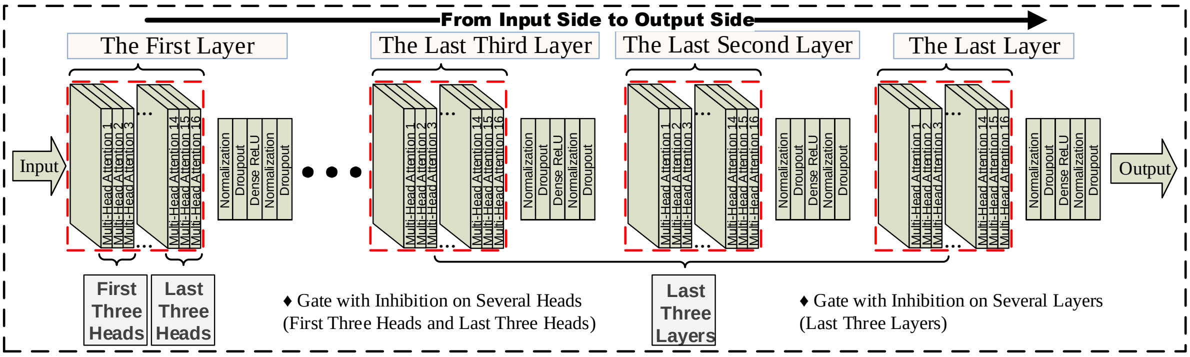


Figure 5.4: Roughly disassembled DeBERTaV3 architecture.

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

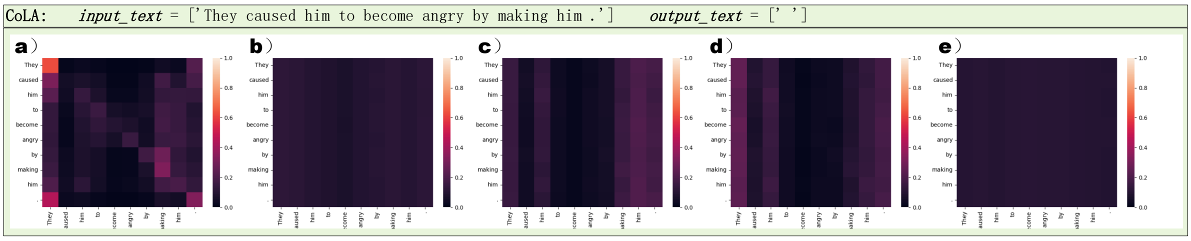


Figure 5.5: From left to right, fine-tuning $BERT - large$ on CoLA with **a)** no-InA, **b)** InA(0.0), **c)** InA(0.1), **d)** InA(0.3), **d)** InA(0.9).

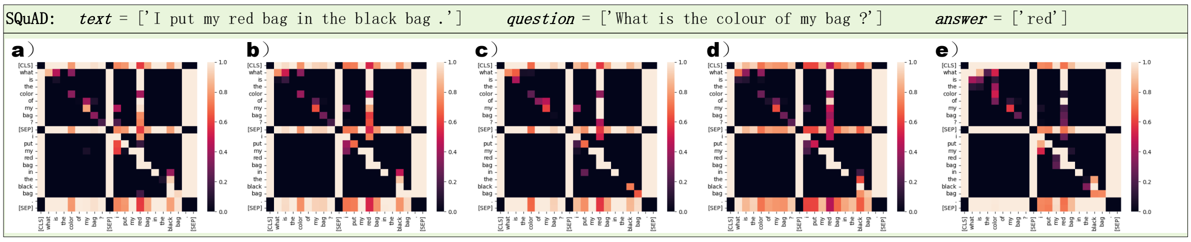


Figure 5.6: From left to right, fine-tuning $BERT - large$ on SQuAD with **a)** no-InA, **b)** InA(0.0), **c)** InA(0.1), **d)** InA(0.3), **e)** InA(0.9).

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) [245]) 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 [246]. This finally prompts RoBERTa to gain a robust speciality - the capacity to handle large and wide vocabularies [220]. InA on

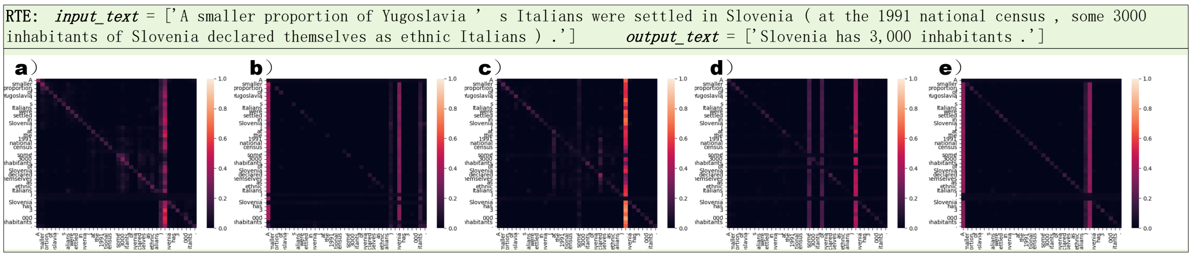

RTE: *input_text* = ['A smaller proportion of Yugoslavia ' s Italians were settled in Slovenia (at the 1991 national census , some 3000 inhabitants of Slovenia declared themselves as ethnic Italians) .'] *output_text* = ['Slovenia has 3,000 inhabitants .']

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).

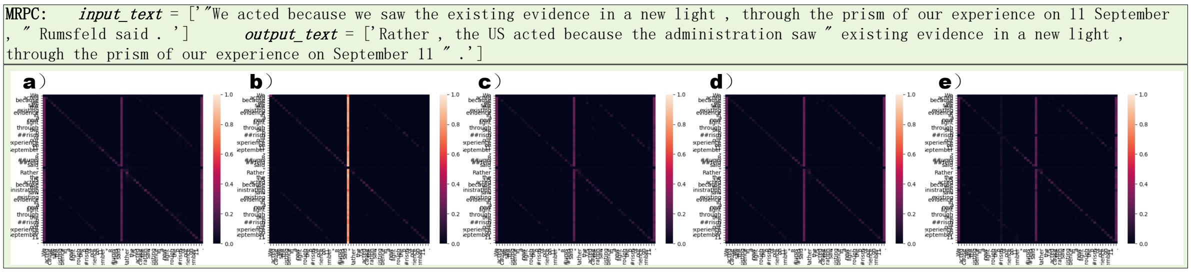

MRPC: *input_text* = ['"We acted because we saw the existing evidence in a new light , through the prism of our experience on 11 September , " Rumsfeld said . '] *output_text* = ['Rather , the US acted because the administration saw " existing evidence in a new light , through the prism of our experience on September 11 " .']

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 [223], 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

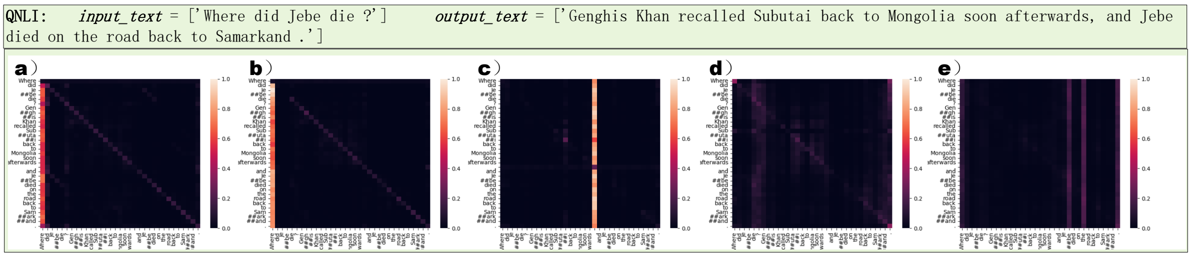
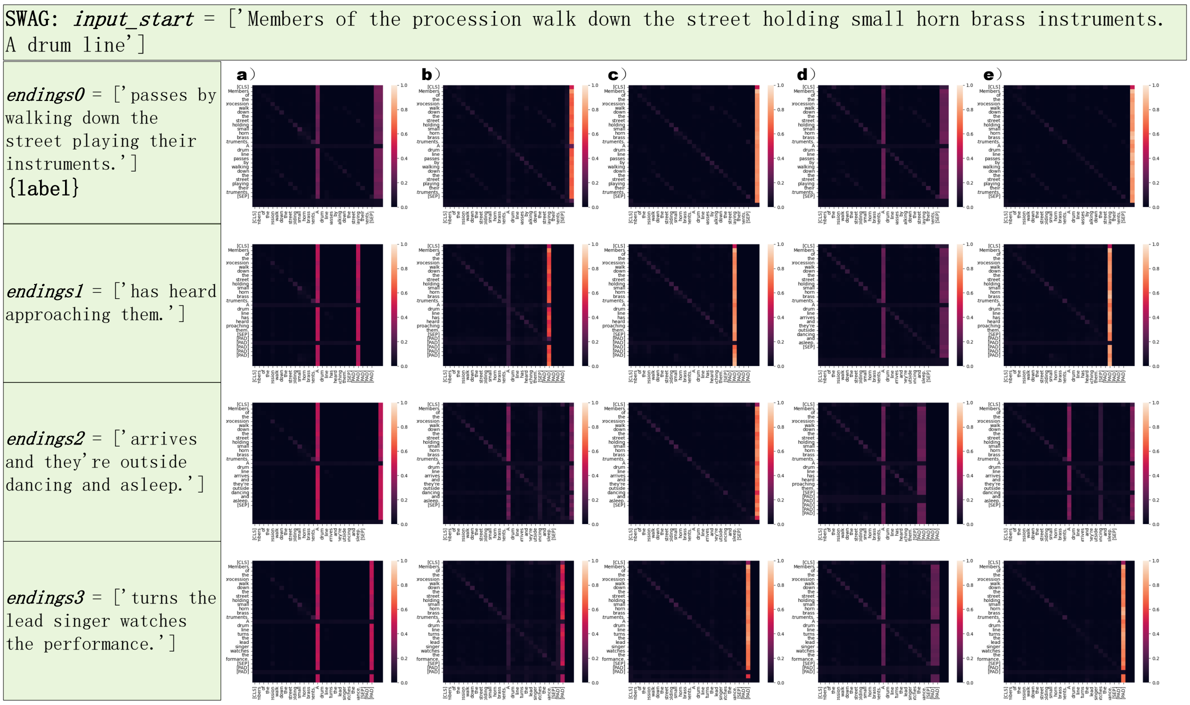

QNLI: *input_text* = ['Where did Jebe die ?'] *output_text* = ['Genghis Khan recalled Subutai back to Mongolia soon afterwards, and Jebe died on the road back to Samarkand .']

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).


SWAG: *input_start* = ['Members of the procession walk down the street holding small horn brass instruments. A drum line']

endings0 = ['passes by walking down the street playing their instruments.'] {label}

endings1 = ['has heard approaching them.']

endings2 = ['arrives and they're outside dancing and asleep.']

endings3 = ['turns the lead singer watches the performance.']

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 [56], chain-of-thoughts reasoning [57], and biomedical diagnosing [58]. Instruction-tuning of LLMs has enabled them to follow natural language instructions and perform real-world tasks [59]. 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 [60], and **(2)** supervised fine-tuning using public benchmarks and datasets augmented with manually or automatically generated instructions [61]. RLHF has proven

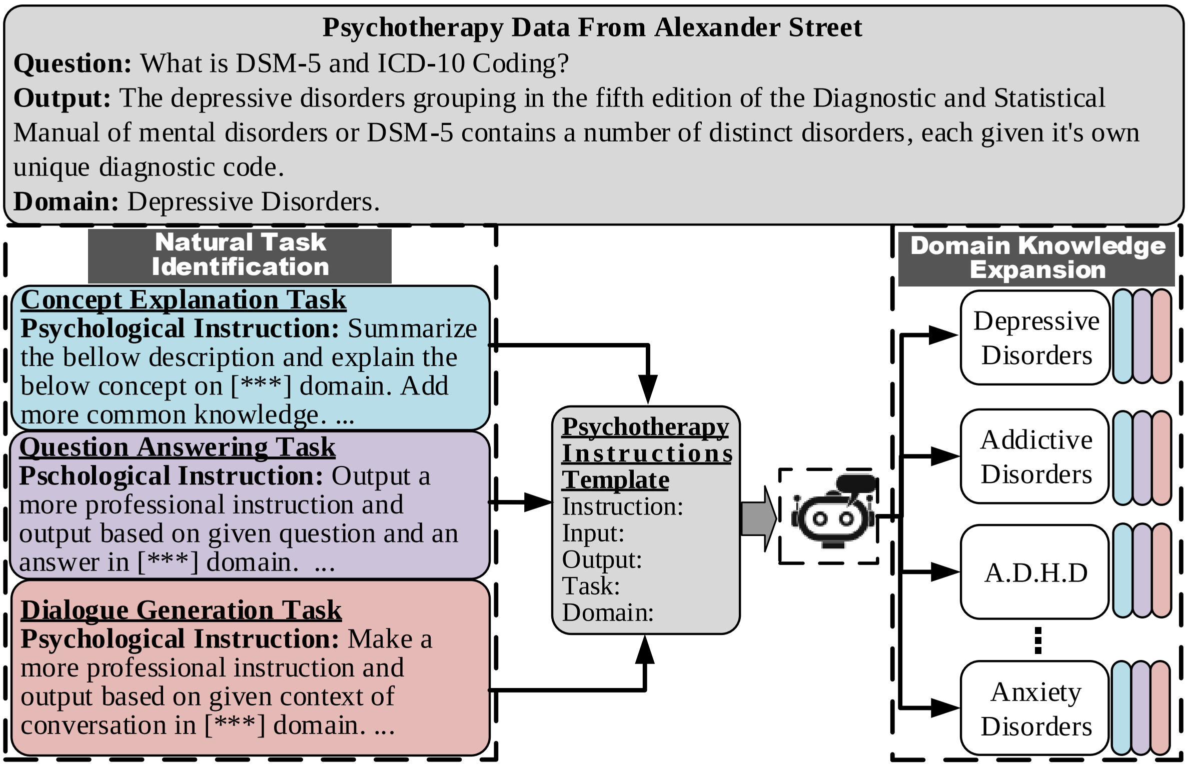

Psychotherapy Data From Alexander Street
Question: What is DSM-5 and ICD-10 Coding?
Output: The depressive disorders grouping in the fifth edition of the Diagnostic and Statistical Manual of mental disorders or DSM-5 contains a number of distinct disorders, each given it's own unique diagnostic code.
Domain: Depressive Disorders.
Natural Task Identification
Concept Explanation Task
Psychological Instruction: Summarize the bellow description and explain the below concept on [***] domain. Add more common knowledge. ...
Question Answering Task
Pschological Instruction: Output a more professional instruction and output based on given question and an answer in [***] domain. ...
Dialogue Generation Task
Psychological Instruction: Make a more professional instruction and output based on given context of conversation in [***] domain. ...
Psychotherapy Instructions Template
Instruction:
Input:
Output:
Task:
Domain:
Domain Knowledge Expansion
Depressive Disorders
Addictive Disorders
A.D.H.D
Anxiety Disorders

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 [62], knowledge graphs [63], and biomedical applications [64], but it comes with a high cost. Natural instructions [59], and even un-natural instructions [65], can provide knowledge in multiple domains, but LLMs pre-trained on vast corpora (e.g., Llama1 [66] and Llama2 [67] containing books, common crawled conversations, arxiv articles, GitHub, C4, and Wikipedia data) still require additional professional knowledge, especially from domain experts. Self-Instruct tuning [68], [69] 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 [247], teaching small language models to reason [248], adapting LLMs on natural common domains through Self-Instruct tuning with low cost [68], Stanford Alpaca [249] using 52K instruction-following samples generated by GPT-4, and Vicuna [250] relying on around 700K instruction-following samples (70K conversions) shared with user-ChatGPT [251]. 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 [252]–[254]. 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: $M\left(I_t, X_{t,i}\right) = Y_{t,i}$, 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 [255], [256]. Some studies have focused on using smart conversational agents to detect neuropsychiatric disorders [257], [258], employing deep neural learning models for generating psychiatric-oriented responses. Other research [259] 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 [253].

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 [260], [261]. 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 [61] offers an annotation-free method for aligning

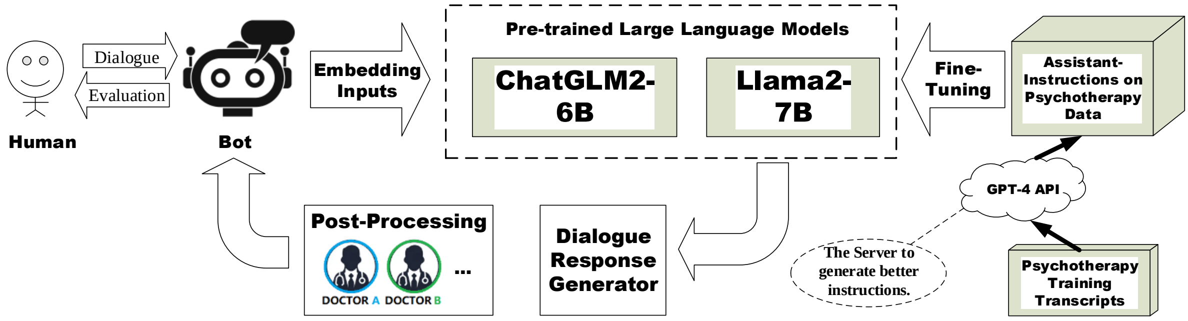

Dialogue
Evaluation
Human
Bot
Embedding Inputs
Pre-trained Large Language Models
ChatGLM2-6B
Llama2-7B
Fine-Tuning
Assistant-Instructions on Psychotherapy Data
GPT-4 API
The Server to generate better instructions.
Psychotherapy Training Transcripts
Post-Processing
DOCTOR A
DOCTOR B
...
Dialogue Response Generator

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 [65] 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 [55], Prefix [52], LoRA [51], GLoRA [231] 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?	**Instruction:** Kindly provide professional suggestions or comments on effectively addressing and alleviating [Depressive Disorders].
Input: We are talking about [Depressive Disorders].	**Input:** We are discussing [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.	**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 [262], [263] 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		Psycho -Eval
				$ROUGH_L$	*LMentry*	
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 [264], Falcon-7B [265], Llama2-7B [67], ChatGLM2-6B [266] and MPT-7B [267] 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 [61].

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$_L$ [268] and LMentry [269] 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? ”””	**Instruction:** ””” What suggestions or comments do you have to address or alleviate [Addiction]? ”””
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. ”””	**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: ””” 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. ”””	**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. ”””	**Task:** ””” Dialogue generation task. ”””
Domain: ””” Addiction. ”””	**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 [61] 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 [270] 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 [59], [60], [65]. 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 [271], 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 images. When the ordinal sequence of the 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 three 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. **Kang, C.***, Guo, Z.*, Chen, Y., Zhang, Y. (2022). Unipolar and Bipolar Disorders Classification: Facial-Based EEGs and DCNN.

2. Wang, H., Guo, Z., Wen, J., Novak, D., **Kang, C.** (2023). Abnormal Brain EEG Networks in Bipolar Disorders under Facial-Emotional Experiments.

3. **Kang, C.**, Prokop, J., Tong, L., Zhou, H., Novak, D. (2023). Gate and Inhibition Mechanism in MLPs and BERTs.

4. Yao, X.*, **Kang, C.***, Zhang, X., Wang, S., Zhang, Y. (2023). 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.**, Hu, Y., Novak, D. (2023). Quantized Embedding for Controllable Diffusion Language Models.

7. **Kang, C.**, Yao, X., Prokop, J., Tong, L., Zhou, H., Hu, Y., Novak, D. (2023). 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] G. Tutz, “Ordinal regression: A review and a taxonomy of models”, *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 14, no. 2, e1545, 2022.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

[49] 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.

[50] 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.

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

[52] 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`.

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

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

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

[60] 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.

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

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

[68] 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.

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

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

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

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

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

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] A. Saeedi, M. Saeedi, A. Maghsoudi, and A. Shalbaf, "Major depressive disorder diagnosis based on effective connectivity in eeg signals: A convolutional neural network and long short-term memory approach", *Cognitive Neurodynamics*, pp. 1–14, 2020, ISSN: 1871-4099.

[89] Z. Wan, J. Huang, H. Zhang, H. Zhou, J. Yang, and N. Zhong, "Hybrideegnet: A convolutional neural network for eeg feature learning and depression discrimination", *IEEE Access*, vol. 8, pp. 30 332–30 342, 2020, ISSN: 2169-3536.

[90] D. Shah, G. Y. Wang, M. Doborjeh, Z. Doborjeh, and N. Kasabov, "Deep learning of eeg data in the neucube brain-inspired spiking neural network architecture for a better understanding of depression", in *International Conference on Neural Information Processing*, Springer, pp. 195–206.

[91] 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.

[92] 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`.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

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

[106] 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.

[107] 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.

[108] 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.

[109] 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`.

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

[111] 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.

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] 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.

[117] 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.

[118] 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.

[119] 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.

[120] 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.

[121] 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.

[122] 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.

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

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

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

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

[127] 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.

[128] 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.

[129] 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.

[130] 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.

[131] 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.

[132] 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.

[133] 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.

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

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

[136] 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.

[137] 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.

[138] 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.

[139] 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.

[140] 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.

[141] 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.

[142] 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.

[143] 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.

[144] 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.

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

[146] 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.

[147] 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.

[148] 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.

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

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

[151] 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.

[152] 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.

[153] 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.

[154] 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.

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

[156] 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.

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

[158] 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.

[159] 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.

[160] 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.

[161] 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.

[162] 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.

[163] 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.

[164] 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.

[165] 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.

[166] 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.

[167] 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.

[168] 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.

[169] 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.

[170] 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.

[171] 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.

[172] 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.

[173] 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.

[174] 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.

[175] 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.

[176] 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`.

[177] 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`.

[178] 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.

[179] 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.

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

[181] 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`.

[182] 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`.

[183] 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`.

[184] 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`.

[185] 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`.

[186] 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`.

[187] 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`.

[188] 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`.

[189] 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`.

[190] 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`.

[191] 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`.

[192] 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`.

[193] 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`.

[194] 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`.

[195] 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`.

[196] 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`.

[197] 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`.

[198] 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`.

[199] 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`.

[200] 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`.

[201] 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`.

[202] 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`.

[203] 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`.

[204] 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.

[205] 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.

[206] 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.

[207] 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.

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

[209] 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.

[210] 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.

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

[212] D. Hendrycks and K. Gimpel, “Bridging nonlinearities and stochastic regularizers with gaussian error linear units”, 2016.

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

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

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

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

[217] 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.

[218] 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.

[219] 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.

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

[221] 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.

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

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

[224] 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.

[225] 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.

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

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

[228] 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.

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

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

[231] 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.

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

[233] 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.

[234] 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.

[235] 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.

[236] 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.

[237] 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.

[238] 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`.

[239] 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.

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

[241] 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.

[242] 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.

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

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

[245] 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`.

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

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

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

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

[250] 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.

[251] ShareGPT. “Sharegpt, https://sharegpt.com/”. (2023), [Online]. Available: `https://sharegpt.com/`.

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

[253] 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.

[254] 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.

[255] 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.

[256] K. T. Pham, A. Nabizadeh, and S. Selek, “Artificial intelligence and chatbots in psychiatry”, *Psychiatric Quarterly*, vol. 93, no. 1, pp. 249–253, 2022.

[257] 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.

[258] 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.

[259] 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.

[260] O. Honovich, U. Shaham, S. R. Bowman, and O. Levy, "Instruction induction: From few examples to natural language task descriptions", *arXiv:2205.10782*, 2022.

[261] 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.

[262] T. Schick and H. Schütze, "Generating datasets with pretrained language models", *arXiv preprint arXiv:2104.07540*, 2021.

[263] V. Sanh, A. Webson, C. Raffel, *et al.*, "Multitask prompted training enables zero-shot task generalization", *arXiv preprint arXiv:2110.08207*, 2021.

[264] OpenAI. "Chatgpt, https://openai.com/blog/chatgpt/". (2023), [Online]. Available: `https://openai.com/blog/chatgpt/`.

[265] 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`.

[266] 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.

[267] 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).

[268] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries", in *Text summarization branches out*, 2004, pp. 74–81.

[269] A. Efrat, O. Honovich, and O. Levy, "Lmentry: A language model benchmark of elementary language tasks", *arXiv preprint arXiv:2211.02069*, 2022.

[270] 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`.

[271] H. W. Chung, L. Hou, S. Longpre, *et al.*, "Scaling instruction-finetuned language models", *arXiv preprint arXiv:2210.11416*, 2022.