EASRL-MH

西北工业大学附属中学
The Middle School Attached To Northwestern Polytechnical University

2025中国国际大学生创新创业大赛：
萌芽赛道

EASRL-MH：基于强化学习的智能精神健康诊断系统

—— DDQN-GNN新时代“家用心理健康体温计”

刘沛卓 | 项目负责人

西北工业大学附属中学高2026届B7班

2025年7月

中国精神健康危机：供需矛盾触目惊心

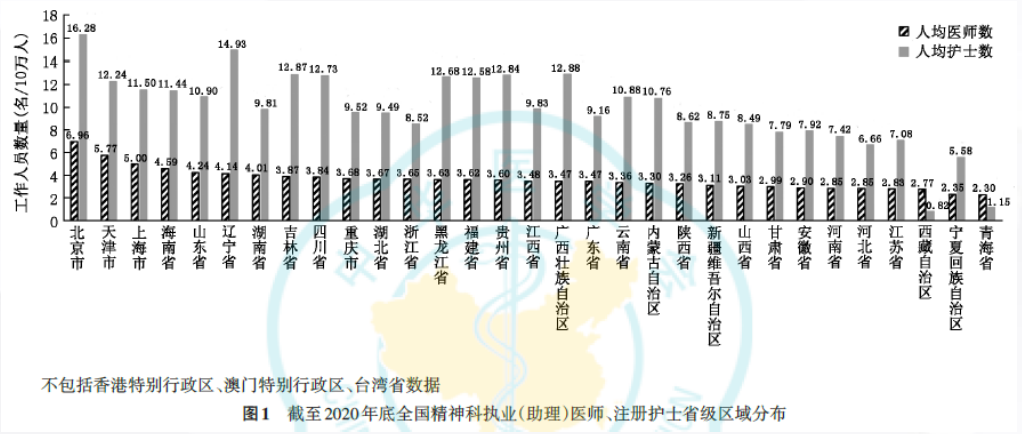
Project Origin

约1.73亿
精神障碍患者数

约5万
精神健康医生数

3.55名/10万人
医生—人口比
远低于WHO建议：10-15 名 / 10 万人

91.2%
从未接受专业治疗的患者比例



不包括香港特别行政区、澳门特别行政区、台湾省数据

图1 截至2020年底全国精神科执业（助理）医师、注册护士省级区域分布

我国精神疾病诊断设施地域分布**极为不均**

全国350个区县(12.31%)无精神卫生医疗机构，
883个区县(31.05%)无精神科床位。
西部地区单位土地面积上精神科开放床位、医师和护士数量较中部相差4倍左右、较东部差7~11倍。

我国每年因**精神疾病**造成的经济损失超过**1万亿元**

数据来源：
1. 马宁, 陈润滋, 张五芳, 等. 2020年中国精神卫生资源状况分析[J]. 中华精神科杂志, 2022, 55(6): 459-468. DOI: 10.3760/cma.j.cn113661-20220617-00158.
2. Huang Y et al. Prevalence of mental disorders in China: a cross-sectional epidemiological study. Lancet Psychiatry. 2019 Mar;6(3):211-224. DOI: 10.1016/S2215-0366(18)30511-X. Epub 2019 Feb 18. Erratum in: Lancet Psychiatry. 2019 Apr;6(4):e11. DOI: 10.1016/S2215-0366(19)30074-4. PMID: 30792114.

国家日益关注精神健康问题

Project Origin

2012-2013年

《中华人民共和国精神卫生法》

1

2

2015-2016年

《全国精神卫生工作规划》

《关于加强心理健康服务的指导意见》

2018-2019年

《全国社会心理服务体系建设试点工作方案》

《健康中国行动》之“心理健康促进行动”

3

4

2022-2024年

《"十四五"国民健康规划》及《国务院关于精神卫生工作情况的报告》

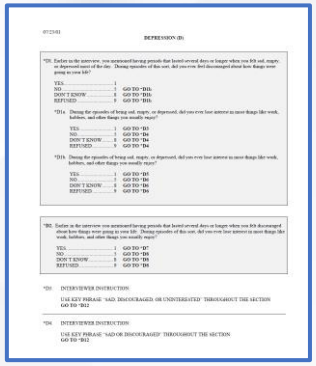
——《精神卫生工作情况报告》强调

“加大科学研究工作力度...利用大数据和人工智能，开展防治一体化技术研发及推广应用。”

现存诊断方法问题
Current Problems

西北工业大学附属中学
The Middle School Attached To Northwestern Polytechnical University

传统诊断模式无法作为日常精神健康筛查方法



效率极低

- 完整CIDI访谈需 300-400 个问题
- 需要专业受训的访谈员控制流程
- 单次诊断耗时 2-3 小时

CIDI(Composite International Diagnostic Interview)
仅Depression模块便有100题左右

诊断成本高昂

- 单次门诊诊断花费约 200~300元
- CIDI/SCID等访谈耗费 近千元
- 且多数消费不纳入医保

*根据陕西省医疗服务项目价格（2024版），单次A类量表价格为22元，临床诊断通常需要数十份量表

主观性强、准确率低

场景	研究	主要方法	敏感度 (≈检出率)	特异度 (≈排除率)
初级保健（全科/内科门诊）—不借助量表，医生日常问诊	Mitchell A J 等 19 项研究荟萃 (The Lancet 2009) [1]	与结构化访谈比对	≈50 %	≈81 %
精神卫生专业人员—使用半结构化面谈（SCID 等）	系统综述 (JAMA 2002) [2]	精神科 / 心理师 vs SCID	κ = 0.64–0.93	—

可及性差、资源紧张

全国350个区县(12.31%)无精神卫生医疗机构，883个区县(31.05%)无精神科床位。

[1] Mitchell AJ, Vaze A, Rao S. “Clinical diagnosis of depression in primary care: a meta-analysis.” The Lancet. 2009 Aug 22;374(9690):609-619. doi:10.1016/S0140-6736(09)60879-5.
[2] Williams JW Jr, Noël PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? JAMA. 2002 Mar 6;287(9):1160-70. doi: 10.1001/jama.287.9.1160. PMID: 11879114.

项目愿景：新时代的“家用心理健康体温计”

Project Background

随着社会和学业压力的不断增长，身边越来越多的人开始呈现出精神健康相关症状。

由于部分地区精神健康诊断资源紧张、精神健康问题早期症状隐蔽，且统一进行大规模精神健康普查的实现难度极高，当患者前往医院就诊时，症状往往已经非常严重。

这种情况可以通过日常心理健康筛查避免，患者通过在家运行心理健康预筛查系统，可以评估自己现在的心理健康状况。

若系统反馈当前风险较高，患者可以选择前往正规医院获取进一步诊断。

然而，现存诊断方法需要在医院进行专业测试和结果解读，且该方法问题数量多、资源紧张，难以满足作为日常筛查方法的要求。

因此，我们的项目希望实现一个：

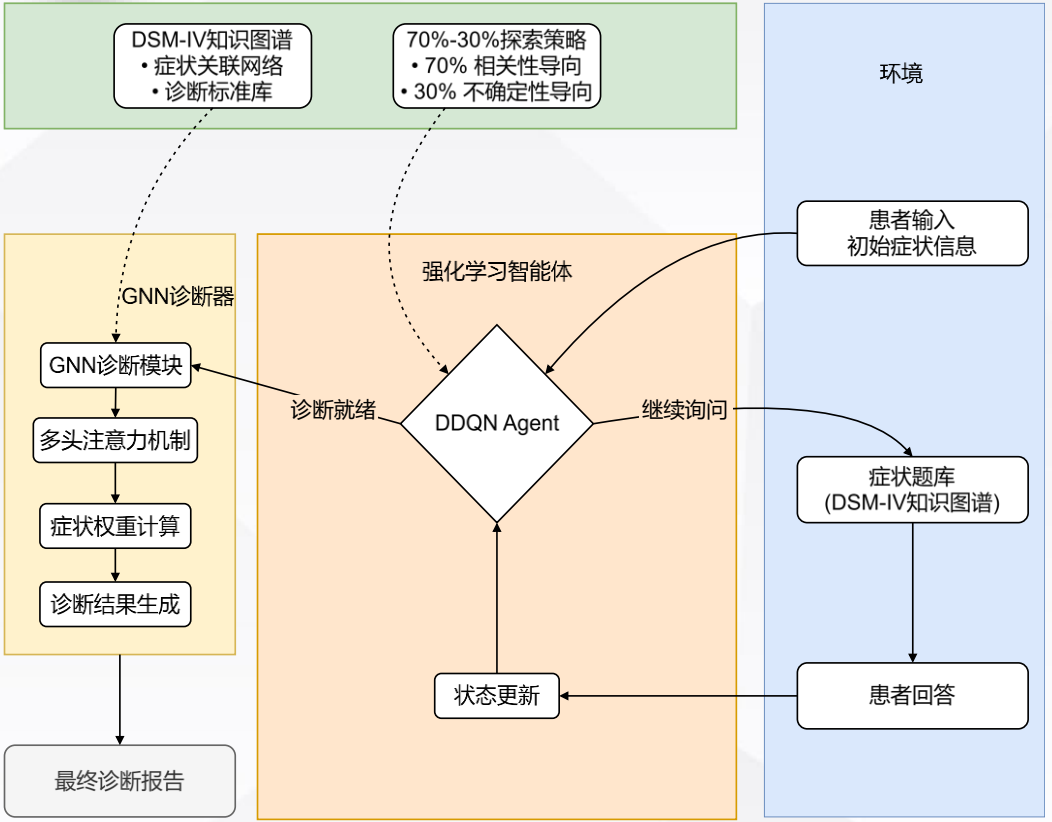
“容易用、问的少、问得准 的心理健康日常筛查方法”

该方法作为一个日常的“心理健康体温计”，可以在患者对自己的心理状况感到不确定时，帮助患者评估自己的心理健康风险程度，决定是否需要前往医院寻求进一步诊断。

基于上述项目愿景，我借用 2024 年英才计划提供的平台，在王柱教授指导下开始研究。研究期间，我主动寻求进一步的专业指导，联系到了西安交通大学的任鹏举教授以及西北大学的张文静老师。并在 2024 年暑假前往加州大学伯克利分校 (University of California, Berkeley) 继续深入学习相关知识。并历时一年左右，提出了我们针对该问题的解决方法：**EASRL-MH.**

本项目创新型实现方案：EASRL-MH

Our Innovative Solution: EASRL-MH



三大亮点：

(相比传统方法，EASRL-MH有什么不同?)

1.相关性-不确定性行为策略

- 我们提出了一种相关性-不确定性引导的 ϵ-greedy 策略
- 智能体根据症状-诊断相关性选择下一步询问
- 相比传统 ϵ-greedy，提高可控性和准确性

2. 症状-诊断二分图神经网络

- 根据症状-诊断原生关系构建诊断器
- 将DSM-IV诊断标准集成到模型中
- 为模型生成诊断提供更客观、可靠的标准

3. 高模型可解释性

- GNN中注意力头可进行可视化，辅助用户理解特征重要性
- 模型训练、询问、决策过程全部可追踪
- 帮助医生、患者理解和相信决策过程

项目优势
Project Advantages

本项目 (EASRL-MH) 具有以下优势：

现存方法缺点：

效率极低

主观性强、准确率低

诊断成本高昂

可及性差、资源紧张

01 效率大幅提升

传统问卷+医生询问 → **自适应智能提问**

- 传统医生提问 -> 40~50个问题
- 传统智能询问策略 ->约15.3次提问
- EASRL-MH -> 约9.6次提问

02 诊断成本低

线下诊断 → **线上低成本预筛查**

- 可本地部署，可离线诊断
- 项目未来预备开源，保证低成本诊断

03 可及性高

- 本项目可用于高风险人群/潜在患者预筛查
- 根据筛查结果，用户可选择是否寻求专业诊断

04 标准客观、准确性高

1. 基于Diagnostic and Statistical Manual (DSM-IV) 进行诊断，保证诊断客观性。

2. 在NCS-R数据集上完成测试，且表现良好

EASRL-MH 系统在 NCS-R 数据集上的性能评估结果

评估指标	EASRL-MH	Greedy Policy (Info Gain)
准确率	**86.4%**	74.7%
精确率	**0.85**	0.72
召回率	**0.82**	0.68
F1 分数	**0.83**	0.70
ROC-AUC	**0.92**	0.79
平均询问症状数	**9.6**	10.1
询问症状标准差	3.9	**2.6**

主要创新点 1：相关性-不确定性引导的策略

Correlation-Exploration Guided Strategy

传统动态方法：纯随机探索(ϵ-greedy策略)

$$\pi(a \mid s) = \begin{cases} \arg\max_a Q(s,a), & \text{prob} = 1-\varepsilon \\ \text{random } a, & \text{prob} = \varepsilon \end{cases}$$

模型在探索(Exploration)时完全随机选择

忽略了诊断和症状间的相关性

传统静态方法：静态特征选择方法

基于 SoftMax 的概率方法

$$P_{\text{static}}(a_i) = \frac{\exp(w(a_i))}{\sum_{j=1}^{n} \exp(w(a_j))}$$

Top-k 硬编码

$$\text{Select } \mathcal{A}_k = \{a_i \in \mathcal{A} \mid w(a_i) \text{ is among top-k}\}$$

模型决策完全依赖于先验知识

无法动态做出决定

本项目提出：

相关性-不确定性引导的ϵ-greedy策略

$$\pi(a \mid s) = \begin{cases} \arg\max_a Q(s,a), & \text{prob} = 1-\varepsilon \\ a \sim P_{\text{knowledge}}(a \mid s), & \text{prob} = \varepsilon \end{cases}$$

$$P_{\text{knowledge}}(a \mid s) = 0.7\, P_{\text{correlation}}(a) \;+\; 0.3\, P_{\text{uncertainty}}(a \mid s)$$

模型在探索(Exploration)步遵循实时相关性引导的探索步骤

技术优势：

1. 基于当前症状和诊断的相关性做出选择
2. 避免盲目的均匀随机探索，提高样本效率
3. 不仅使用症状相关性选择($P_{correlation}$)，同时保证对未知信息的探索($P_{uncertainty}$)

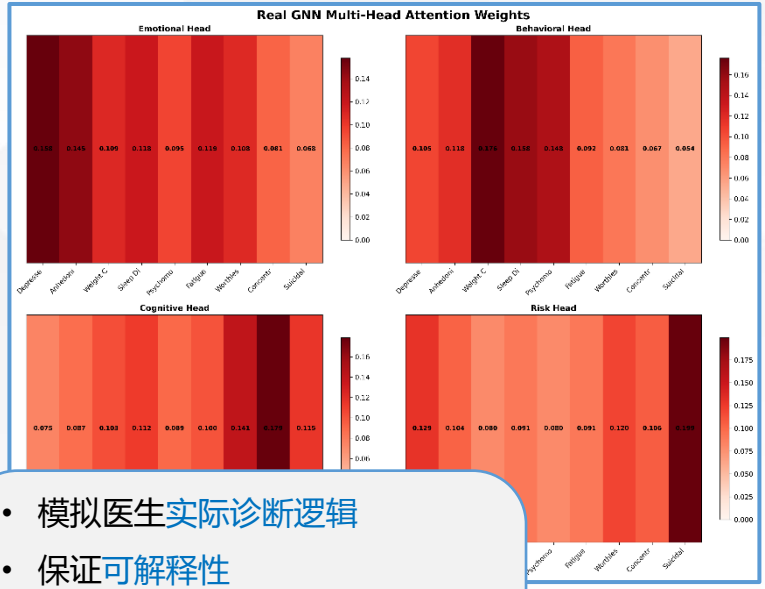
$$P_{\text{correlation}}(a) = \frac{\text{corr}(x_a,\, y)}{\sum_{i=1}^{n} \text{corr}(x_i,\, y)} \qquad P_{\text{uncertainty}}(a \mid s) = \frac{\sum_{d \in D_a} p_d(1-p_d)}{\sum_{i=1}^{n} \sum_{d \in D_i} p_d(1-p_d)}$$

$$\text{corr}(x_a, y) = \frac{\sum_{j=1}^{m} (x_{a,j} - \bar{x}_a)(y_j - \bar{y})}{\sqrt{\sum_{j=1}^{m} (x_{a,j} - \bar{x}_a)^2} \cdot \sqrt{\sum_{j=1}^{m} (y_j - \bar{y})^2}}$$

主要创新点 2：症状-诊断二分图神经网络 (GNN)

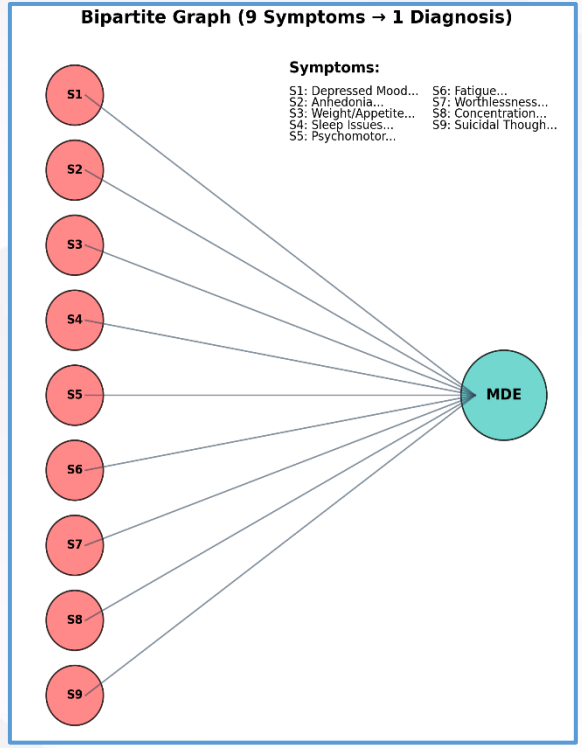
Symptom-Diagnosis Bipartite Graph Neural Network

优势1：多头注意力机制



- 模拟医生实际诊断逻辑
- 保证可解释性
- 提升诊断准确性和鲁棒性
- 动态分配症状权重
- 符合DSM-IV医学标准

优势2：集成DSM-IV症状-诊断标准



Criteria for Major Depressive Episode

A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.

Note: Do not include symptoms that are clearly due to a general medical condition, or mood-incongruent delusions or hallucinations.

(1) depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful). **Note:** In children and adolescents, can be irritable mood.

(2) markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation made by others)

(3) significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day. **Note:** In children, consider failure to make expected weight gains.

DSM-IV 严重抑郁发作诊断标准节选

- 保证诊断的医学权威性
- 为AI学习过程提供医学约束
- 限制模型建立不合理链接

↓

- 避免虚假关联

实验表明：相比传统MLP，EASRL-MH可将诊断准确性提升约15.2%

主要创新点 3：模型具有高可解释性

High Model Interpretability

西北工业大学附属中学
The Middle School Attached To Northwestern Polytechnical University

可解释性
两大核心问题

模型为何选择做出该诊断?

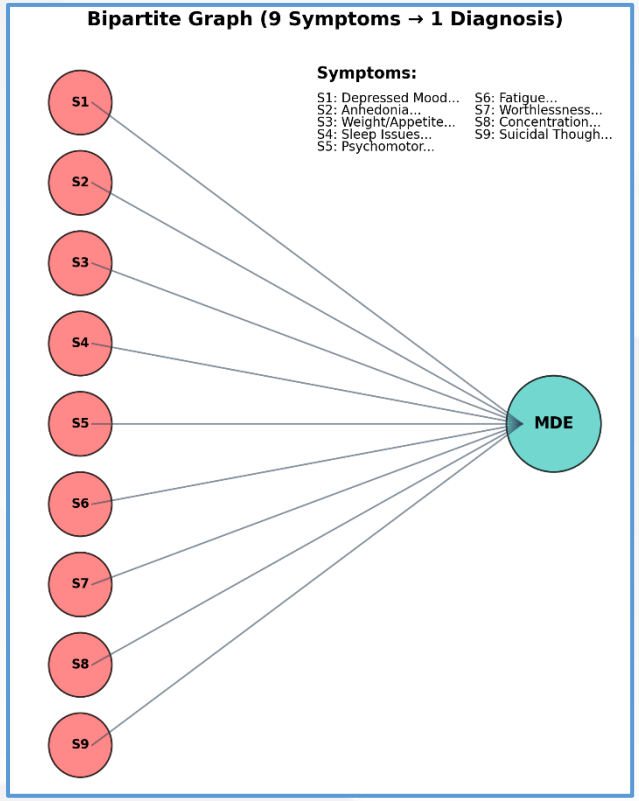
决策模块： 症状-诊断二分图神经网络 (GNN)

1. “基石”：GNN由DSM-IV知识图谱构建
   - 模型诊断基于DSM-IV
   - DSM-IV人类可理解
2. “钥匙”：多头注意力可视化
   - 量化不同症状对该诊断结果的贡献程度

模型为何选择该问题进行提问?

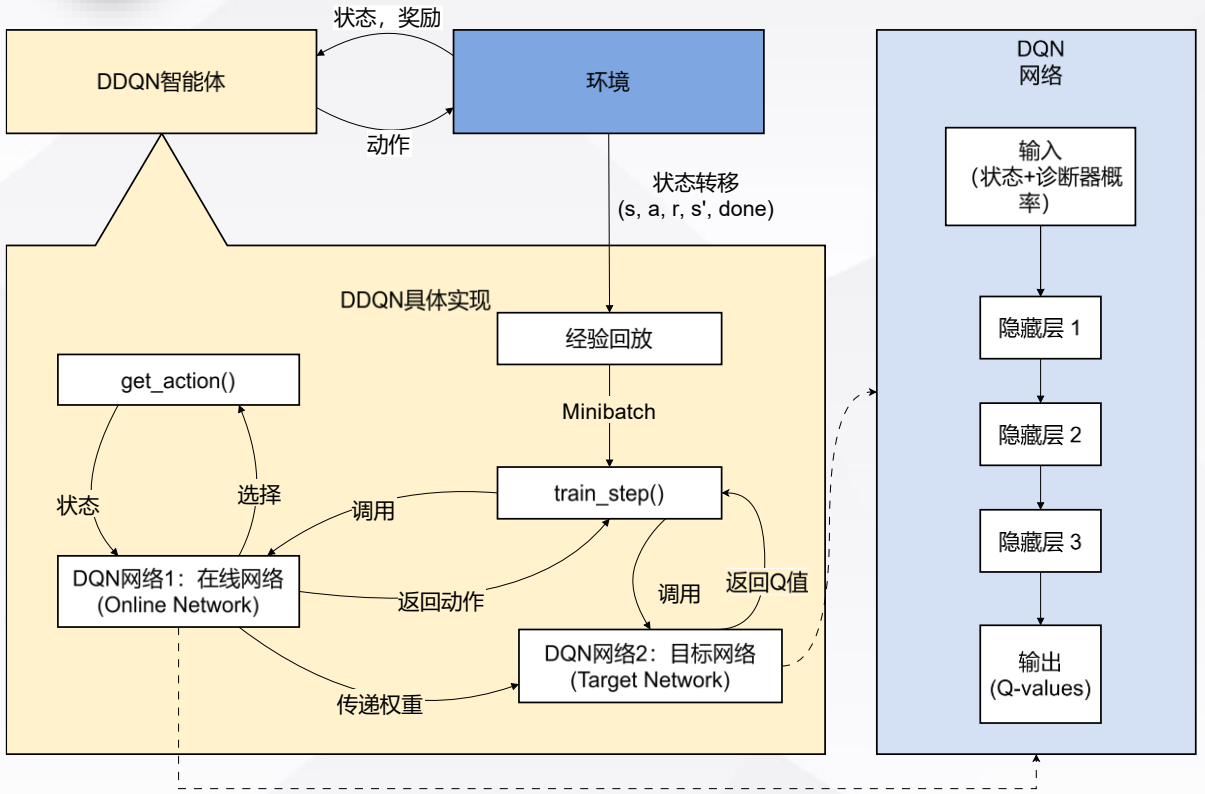
询问模块： 深度Q网络智能体(DDQN)

1. “再现”：智能体动作序列
   - 帮助理解询问过程
   - 医生通过查看询问过程，可选择是否信任结果
2. “量化”：相关性-不确定性引导
   - 决策引导模式由数学公式明确指定
   - 做出该决定：70%可能因为症状相关性，30%可能因为对诊断不确定



关键技术 1：增强双重深度Q-网络 (DDQN)

Double Deep Q-Network Agent



强化学习环境定义为**部分可观测马尔可夫决策过程 (POMDP)**.

选用**双重深度Q-网络 (DDQN)** 作为强化学习环境中的智能体，包含**两个独立子网络**:

在线网络 (Online Network)

- 计算Q-Value
- 选择下一步动作

目标网络 (Target Network)

- 审核Q-Value
- 确保模型不会“过于乐观”

有效缓解单 DQN 高估 Q-Value 问题

基线测试
Baseline Comparison

我们在本地对现存主流智能诊断方法进行复现，并在 (National Comorbidity Survey – Replication, NCS-R) 数据集上进行基线测试，该数据集采用了世界卫生组织（WHO）开发的 Composite International Diagnostic Interview (CIDI) 工具，按照 DSM-IV 标准评估精神疾病。
实验结果如表1所示：

EASRL-MH全面刷新了在MDE诊断上的 (State-of-the-art, SOTA) 水平。

在所有关键指标上均达到最佳性能

- 准确率: **86.4%**
  (相比次优XGBoost提升约4.0%)
- F1分数: **83.4%**
  (相比次优MLP提升约4.4%)
- AUC-ROC: **91.8%**
  (相比次优MLP提升约3.5%)
- 平均查询数: **9.6次**
  (相比大多方法均大幅提升)

综合评估下，
EASRL-MH方法为**现存最优方法**

表 1: 基准测试结果对比

模型	准确率	F1 分数	AUC-ROC	平均查询数
传统机器学习方法				
Random Forest	0.812 ± 0.018	0.763 ± 0.019	0.856 ± 0.012	-
XGBoost	0.824 ± 0.016	0.776 ± 0.017	0.871 ± 0.014	-
SVM	0.798 ± 0.023	0.749 ± 0.024	0.843 ± 0.016	-
Logistic Regression	0.766 ± 0.021	0.716 ± 0.022	0.819 ± 0.018	-
Decision Tree	0.731 ± 0.032	0.680 ± 0.033	0.784 ± 0.025	-
Symptom Count	0.642 ± 0.019	0.594 ± 0.020	0.695 ± 0.022	-
深度学习方法				
MLP	0.834 ± 0.019	0.790 ± 0.020	0.883 ± 0.013	-
CNN	0.819 ± 0.024	0.769 ± 0.025	0.864 ± 0.017	-
基准策略方法				
Random Policy	0.498 ± 0.031	0.498 ± 0.032	0.521 ± 0.029	14.6 ± 0.8
Greedy Policy (Frequency)	0.721 ± 0.028	0.674 ± 0.029	0.768 ± 0.022	11.2 ± 2.8
Greedy Policy (Info Gain)	0.747 ± 0.024	0.701 ± 0.025	0.792 ± 0.019	10.1 ± 2.6
Fixed Sequence	0.693 ± 0.026	0.646 ± 0.027	0.741 ± 0.021	12.0 ± 0.0
Adaptive Threshold	0.763 ± 0.022	0.717 ± 0.023	0.815 ± 0.017	8.7 ± 2.1
本文提出方法				
EASRL-MH	**0.864 ± 0.015**	**0.834 ± 0.016**	**0.918 ± 0.011**	**9.6 ± 3.9**

注： 表中数值为 5 折交叉验证的平均值 ± 标准差

项目应用
Project Applications

EASRL-MH
心理健康诊断
辅助系统

赋能临床诊断

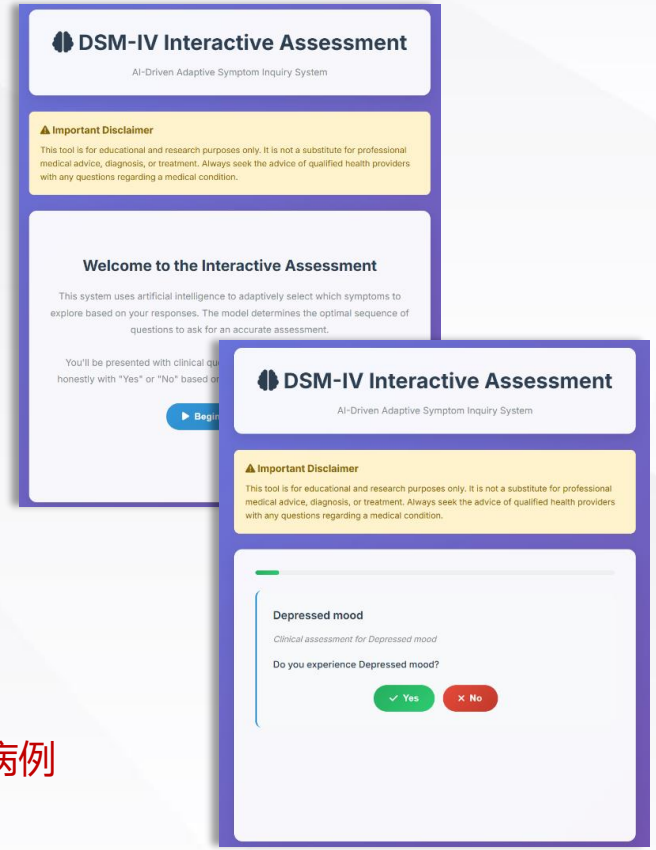
- 作为医生的“智能助手”辅助诊断
- 提供标准化预筛查
- 减少主观性和漏诊

普及公共卫生服务

- 提供大规模心理健康筛查
- 偏远地区提供初步诊断支持
- 缓解地域资源不均

优化医疗资源配置

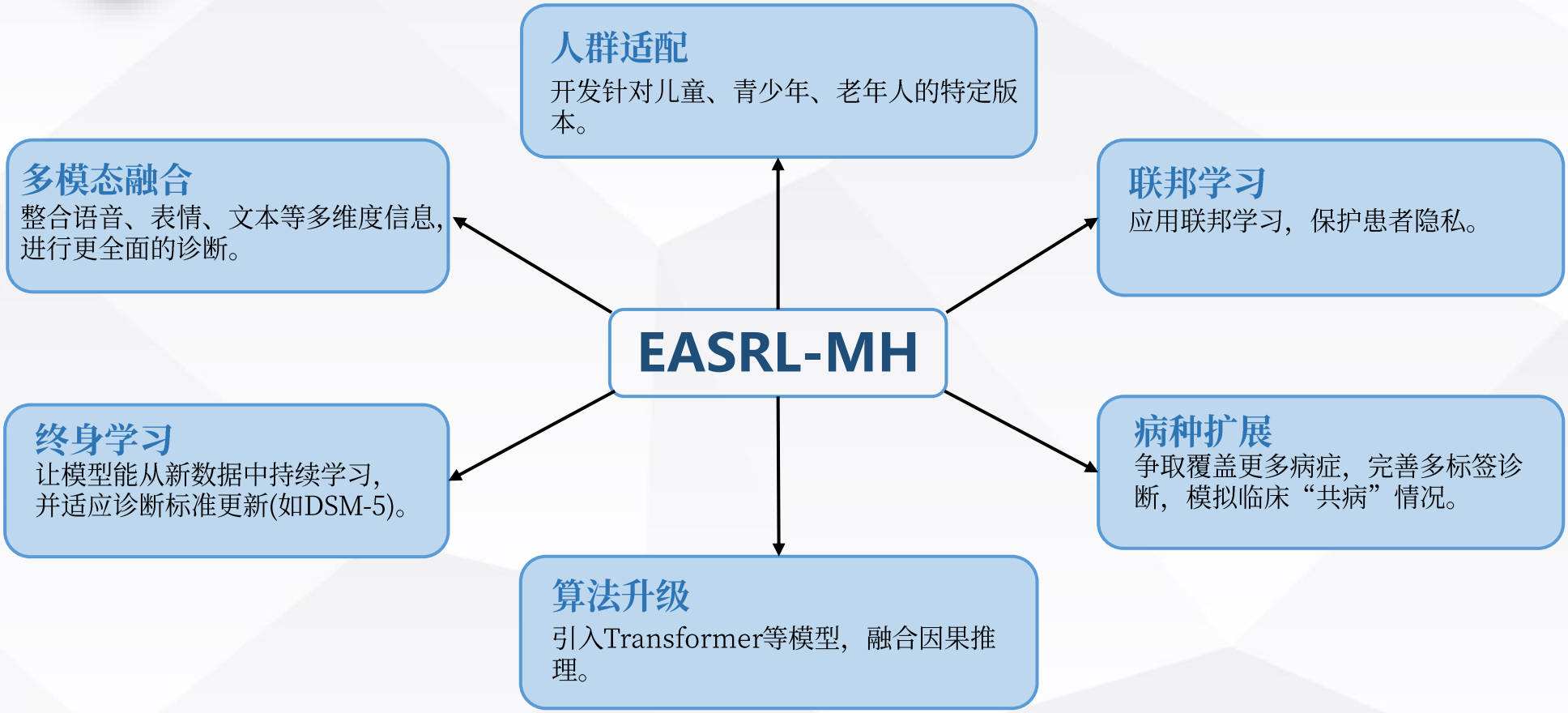
- 实现智能分诊，让专家资源集中于复杂病例
- 提升精神健康服务的可及性



EASRL-MH诊断界面

未来工作
Future Work

我们提出了以下几项潜在改进：



项目团队
Our team

西北工业大学附属中学
The Middle School Attached To Northwestern Polytechnical University

刘沛卓（项目负责人）

西北工业大学附属中学高2026届B7班

负责项目的代码编写、论文撰写等，并承担本项目的负责人。

学术经历：

美国数学竞赛 (AMC12) 全球 Top 5%

加州大学伯克利分校 (UC Berkeley) 夏校，两门大学课程取得A+/A

ACSL 个人满分奖，团队金奖 (队长)

2025 春季 Pioneer Academics 学员

2025 斯坦福大学 (Stanford University) 夏校录取

美国数学邀请赛 (AIME) 9分

美国计算机奥林匹克大赛 (USACO) 铂金组

2024 年英才计划学员，西北工业大学於志文、王柱教授指导

2024 英才计划全国报告陕西省唯一入选学员

2024 英才计划冬令营、夏令营优秀学员等

姜凯然

西北工业大学附属中学高2026届B7班

负责项目实验和测试、实验数据分析和可视化等工作。

学术经历：

区级三好学生

优秀团员

优秀班干部

两岸国际书画大赛特等奖

全国中学生教学建模能力大赛一等奖

全国中学生创新作文大赛初赛一等奖

国画山水九级等

指导老师
Instructors

西北工业大学附属中学
The Middle School Attached To Northwestern Polytechnical University

任鹏举教授

- 西安交通大学人工智能与机器人研究所副所长
- 国家重点研发计划（重点专项）首席科学家
- 国家级人才
- **提供项目整体技术指导，确保技术可行性与先进性**

张文静老师

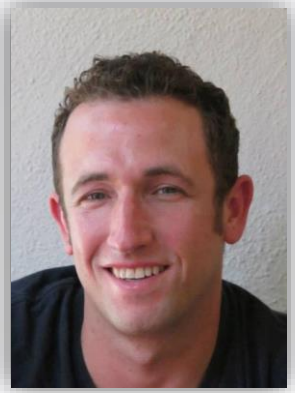
- 西北大学心理中心
- 中国心理卫生学会精神分析专业委员会青年委员
- 在英中国教育研究会CERA成员
- **提供精神健康领域的专业支持，审阅心理学标准匹配性**

贺华老师

- 西北工业大学资产公司（国家大学科技园）副总经理
- 全国优秀创新创业导师
- **指导项目发展方向与应用规划，推动成果转化可行性及外部合作路径设计**

王柱教授

- 西北工业大学计算机学院
- 多项国际知名学术会议最佳论文奖
- **英才计划导师**
- **在英才计划阶段提供AI技术核心支持**



Prof. Rothman

- UC Berkeley, Adjunct Instructor
- University of San Francisco, Adjunct Professor
- **UC Berkeley夏校授课老师**
- **在夏校期间提供模型思路启发**

项目评估
Project evaluation

西北工业大学附属中学
The Middle School Attached To Northwestern Polytechnical University

任鹏举教授

- 西安交通大学人工智能与机器人研究所副所长、教授
- 国家重点研发计划（重点专项）首席科学家（2022年）
- 中国人工智能产业发展联盟 芯片组秘书长
- 工业和信息化部人工智能关键技术和应用评测 学术委员会委员
- 于 2012 年与 MIT 合作发布“HORNET: A Cycle-Level Multicore Simulator”；于 2024 HPCA 发表 Differential-Matching Prefetcher for Indirect Memory Access"；主编国家级教材《人工智能工程技术人员（芯片产品实现）》；主持起草中国通信行业标准《人工智能芯片基准测试评估方法（YD/T3944-2021）》。

EASRL-MH围绕“智能化精神健康筛查”的核心需求展开，旨在结合人工智能与精神健康领域的医学标准，为普通民众提供一种便捷、低成本且高效的心理健康预筛查系统。在当前精神健康资源紧张的背景下，该项目具有高社会应用价值，且能够有效弥补现有传统诊断方法的局限性。

- **创新性与技术先进性**

  本项目采用**强化学习**与**图神经网络**等技术，为精神健康诊断提供智能化、数据驱动的解决方案。相比传统问卷调查和临床访谈，其在效率和准确性上的提升具有重要意义。同时通过**相关性-不确定性引导的探索策略**与**症状-诊断二分图神经网络**的结合，显著提高了诊断的精度与可解释性。

- **跨学科协作**

  该项目充分结合了人工智能与心理健康领域的专业知识，确保了诊断过程的医学权威性。团队成员跨学科的协作，体现了项目在人工智能、心理学、医学和数据科学等领域的融合，展示了**多学科协作**在解决复杂社会问题中的巨大潜力。

综上所述，本项目不仅在技术实现上具备创新性和可行性，也在社会实际需求上具有广泛的应用前景。特别是在提升精神健康服务质量、推动AI技术落地等方面，具有较高的实践价值和广泛的社会影响力。希望项目团队能够持续推动项目研发，并探索更多与临床实际结合的应用场景。

项目评估
Project evaluation

西北工业大学附属中学
The Middle School Attached To Northwestern Polytechnical University

张文静老师

- 西北大学心理健康教育中心专职教师，硕导
- 在英中国教育研究会CERA成员
- 中国心理卫生学会精神分析专业委员会青年委员
- 陕西省健康心理学专业委员会副主任委员兼秘书长
- 著有《大数据视域下危机学生心理画像的建模分析与应用》（国家心理健康与精神卫生中心优秀论文）《自然语言处理技术在心理危机干预中的应用研究》“Music-Based Brain-Computer Interfaces for Mental Disorder Diagnosis: A Multimodal Neurocomputational Approach”(2025 WCP) 等

本项目以“构建高效率、高可及性、智能化的精神健康预筛查系统”为目标，选题立意契合国家《**“十四五”国民健康规划**》关于“**加强心理健康服务供给，推动智能化筛查与干预体系建设**”的指导精神，具备突出的现实意义与推广价值。

从心理学专业视角出发，我认为本项目具备以下几点亮点：

- **心理学理论基础扎实，临床标准明晰**

  项目所采用的 DSM-IV 为国际通用的精神障碍诊断标准，能够有效规范症状归因与模型输出，为AI诊断结果提供心理学与医学的双重保障。

- **高效低成本的技术方案，具有普及潜力**

  项目提出的“日常心理健康体温计”理念，结合本地可部署、低资源依赖等优势，特别适合用于家庭日常使用，以及中小学校园、基层社区、偏远地区等精神健康服务薄弱区域，推动心理服务向早筛、早干预前移。

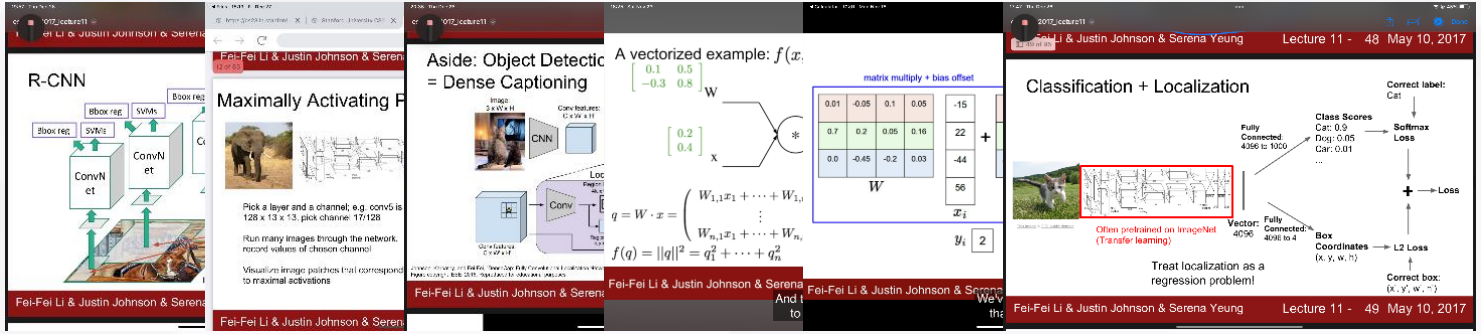
因此，我认为该项目不仅在学术探索上具备创新性，在现实应用中也拥有广阔前景，建议后续继续推进临床合作与真实环境测试，推动科研成果的社会化转化落地。

项目历程
Timeline

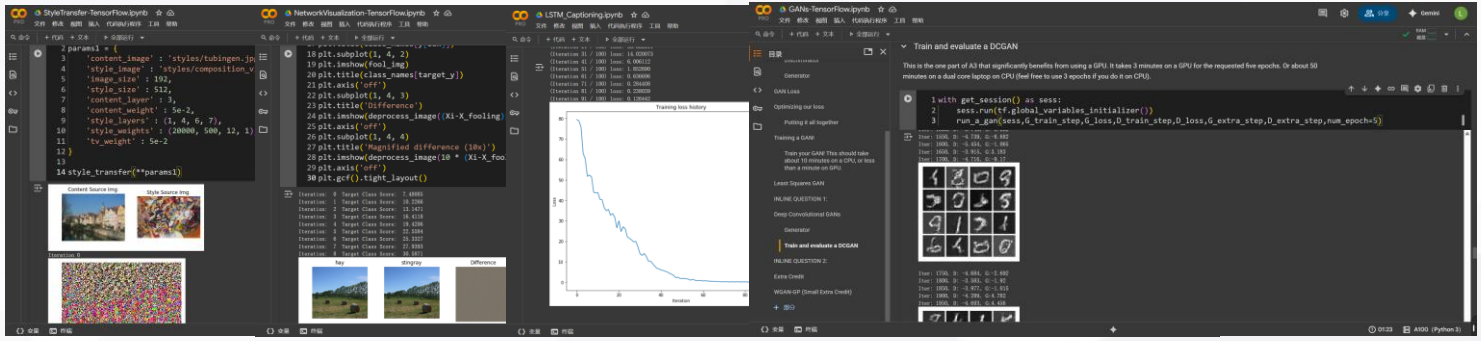
I 知识奠基 (2023)

2023 年：完成任鹏举教授推荐的 Stanford 相关课程，建立AI基础。

Stanford CS231n: Deep Learning for Computer Vision
部分课程剪影



CS231n 部分作业截图



项目历程
Timeline

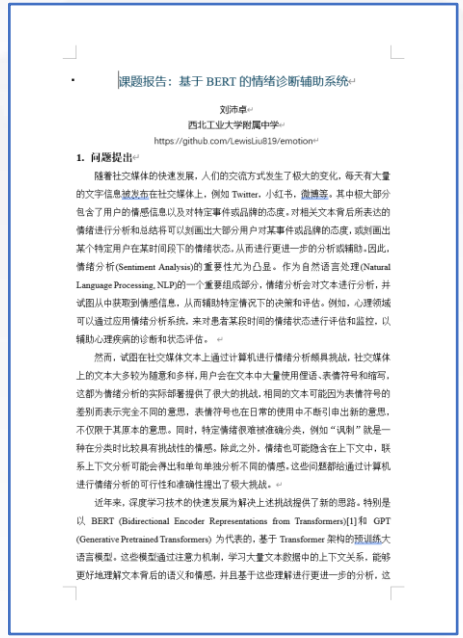
II 项目初步探索与子模块研究 (2024春)

2024 年 1 月：作为"英才计划"项目正式启动




2024 年 1-2 月：确立 “AI驱动的心理健康诊断系统” 研究方向

2024 年 4 月：完成子模块《基于Fine-tuning BERT的情绪诊断辅助系统》



课题报告：基于 BERT 的情绪诊断辅助系统

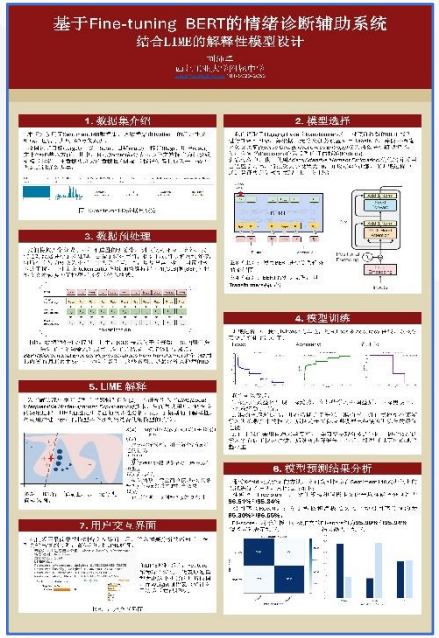
刘沛卓
西北工业大学附属中学
https://github.com/LewisLiu819/emotion

1. 问题提出

随着社交媒体的快速发展，人们的交流方式发生了极大的变化，每天有大量的文字信息被发布在社交媒体上，例如 Twitter、小红书、微博等。其中极大部分包含了用户的情感信息以及对特定事件或品牌的态度。对相关文本背后所表达的情绪进行分析和总结将可以刻画出大部分用户对某事件或品牌的态度，或刻画出某个特定用户在某时间段下的情绪状态。从而进行更进一步的分析或辅助。因此，情绪分析(Sentiment Analysis)的重要性尤为凸显。作为自然语言处理(Natural Language Processing, NLP)的一个重要组成部分，情绪分析会对文本进行分析，并试图从中获取到情感信息，从而辅助特定情况下的决策和评估。例如，心理领域可以通过应用情绪分析系统，来对患者某段时间的情绪状态进行评估和监控，以辅助心理疾病的诊断和状态评估。

然而，试图在社交媒体文本上通过计算机进行情绪分析颇具挑战，社交媒体上的文本大多较为随意和多样，用户会在文本中大量使用俚语、表情符号和缩写，这都为情绪分析的实际部署提供了很大的挑战。相同的文本可能因为表情符号的差别而表示完全不同的意思，表情符号也在日常的使用中不断引申出新的意思，不仅限于其原本的意思。同时，特定情绪很难被准确分类，例如“讽刺”就是一种在分类时比较具有挑战性的情感。除此之外，情绪也可能隐含在上下文中，联系上下文分析可能会得出和单句单独分析不同的情感。这些问题都给通过计算机进行情绪分析的可行性和准确性提出了极大挑战。

近年来，深度学习技术的快速发展为解决上述挑战提供了新的思路。特别是以 BERT (Bidirectional Encoder Representations from Transformers)[1]和 GPT (Generative Pretrained Transformers) 为代表的，基于 Transformer 架构的预训练大语言模型。这些模型通过注意力机制，学习大量文本数据中的上下文关系，能够更好地理解文本背后的语义和情感，并且基于这些理解进行更进一步的分析，这



《基于Fine-tuning BERT的情绪诊断辅助系统》部分项目文件

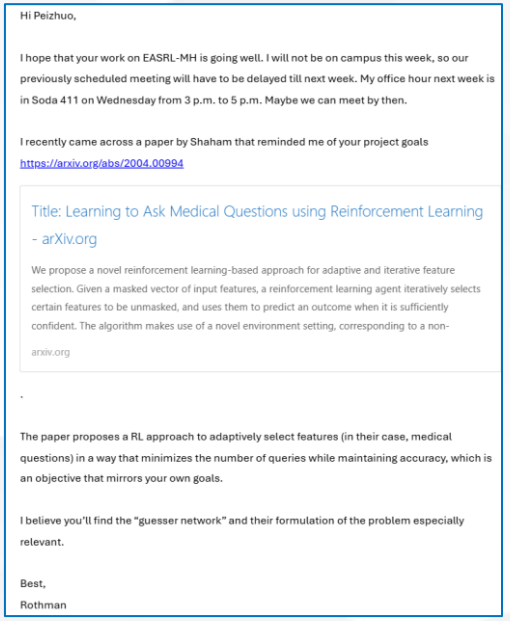
项目历程

Timeline

III 核心技术攻坚 (2024夏-2025春)

2024年6月：加州大学伯克利分校 (UC Berkeley) 访学，确定项目主题为

基于**强化学习的自适应智能精神健康诊断系统**



Hi Peizhuo,

I hope that your work on EASRL-MH is going well. I will not be on campus this week, so our previously scheduled meeting will have to be delayed till next week. My office hour next week is in Soda 411 on Wednesday from 3 p.m. to 5 p.m. Maybe we can meet by then.

I recently came across a paper by Shaham that reminded me of your project goals

https://arxiv.org/abs/2004.00994

Title: Learning to Ask Medical Questions using Reinforcement Learning - arXiv.org

We propose a novel reinforcement learning-based approach for adaptive and iterative feature selection. Given a masked vector of input features, a reinforcement learning agent iteratively selects certain features to be unmasked, and uses them to predict an outcome when it is sufficiently confident. The algorithm makes use of a novel environment setting, corresponding to a non-

arxiv.org

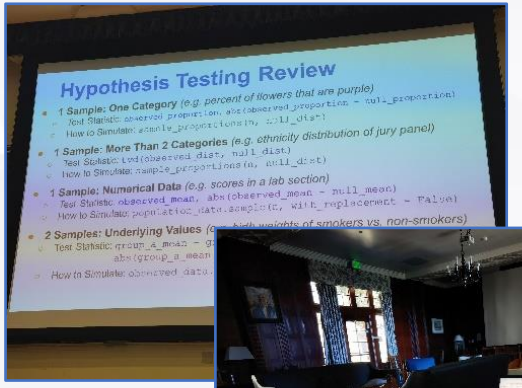
.

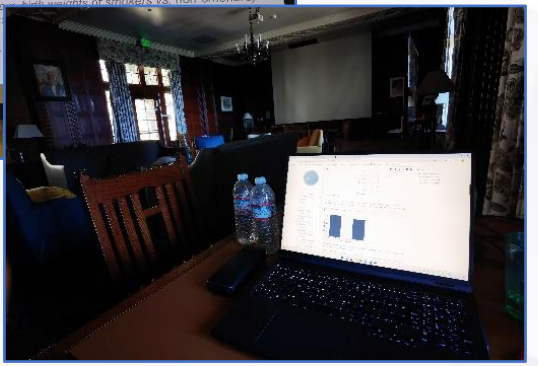
The paper proposes a RL approach to adaptively select features (in their case, medical questions) in a way that minimizes the number of queries while maintaining accuracy, which is an objective that mirrors your own goals.

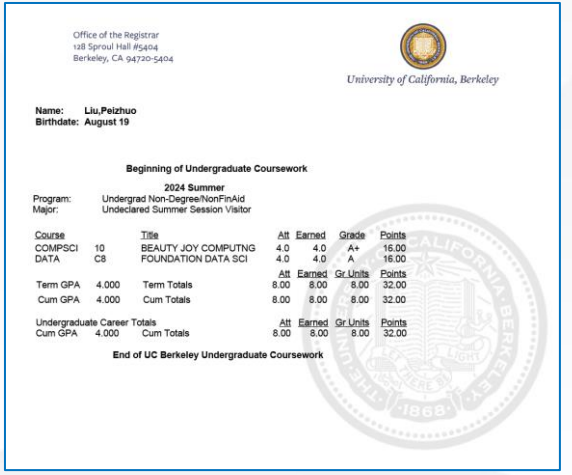
I believe you'll find the "guesser network" and their formulation of the problem especially relevant.

Best,
Rothman

与导师项目交流邮件截图







Office of the Registrar
108 Sproul Hall #5404
Berkeley, CA 94720-5404

University of California, Berkeley

Name: Liu,Peizhuo
Birthdate: August 19

Beginning of Undergraduate Coursework

2024 Summer

Program: Undergrad Non-Degree/NonFinAid
Major: Undeclared Summer Session Visitor

Course		Title	Att	Earned	Grade	Points
COMPSCI	10	BEAUTY JOY COMPUTNG	4.0	4.0	A+	16.00
DATA	C8	FOUNDATION DATA SCI	4.0	4.0	A	16.00

			Att	Earned	Gr Units	Points
Term GPA	4.000	Term Totals	8.00	8.00	8.00	32.00
Cum GPA	4.000	Cum Totals	8.00	8.00	8.00	32.00

			Att	Earned	Gr Units	Points
Undergraduate Career Totals						
Cum GPA	4.000	Cum Totals	8.00	8.00	8.00	32.00

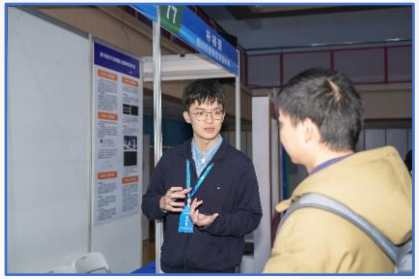
End of UC Berkeley Undergraduate Coursework

项目历程
Timeline

III 核心技术攻坚 (2024夏-2025春)

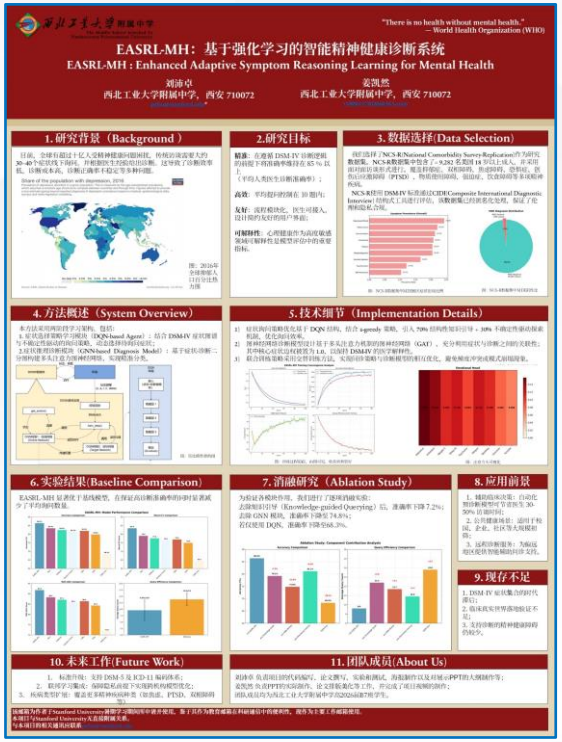
2024年12月：在英才计划全国报告对上述子模块进行汇报

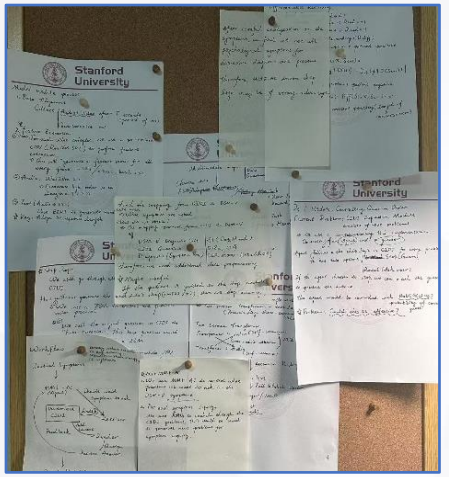
2024年7月-2025年5月：阶段性研发与迭代



于华中科技大学参加 2024 年英才计划全国报告
并汇报子模块《基于Fine-tuning BERT的情绪诊断辅助系统》

部分相关项目文档和笔记展示





Interpretable Artificial Intelligence:
A Survey of Methods and Applications

Peizhuo Liu

High School attatched to Northwestern Polytechnical University

October 17, 2024

Abstract

Interpretable artificial intelligence (XAI) has become a critical area of research in the last five years, driven by the need to understand and trust complex machine learning models. This literature review surveys recent theoretical and methodological developments in interpretable AI methods, with an emphasis on prominent model-agnostic explanation techniques such as SHAP (SHapley Additive exPlanations) [33] and LIME (Local Interpretable Model-Agnostic Explanations) [38]. We outline the mathematical formulations of SHAP and LIME and discuss their key theoretical contributions, including the game theoretic foundations of SHAP and the locally linear surrogate approach of LIME.

Beyond these, we cover a broad spectrum of interpretability methods: intrinsically interpretable models (such as sparse linear models [33], generalized additive models [21], and decision trees [9]), post-hoc explanation methods (encompassing feature attribution, counterfactual explanations [55], rule extraction, and example-based reasoning), visualization-based techniques (notably saliency maps [47] and attention-based explanations for deep networks [7]), and emerging hybrid approaches that integrate interpretability into model design.

We provide an academic analysis of each class of methods, discussing their strengths, limitations, and the trade-offs involved in fidelity, interpretability, and scope (local vs. global). We also compare these approaches to highlight how they complement each other in explaining different aspects of black-box models. The review is structured as a scholarly article with an introduction to the field, detailed sections on each category of methods, and a conclusion summarizing current trends and future directions. All methods are discussed with technical depth and supported by citations from peer-reviewed literature.

未来展望 A glance into the future

我们相信，EASRL-MH 不只是一个技术项目，更是一次精神健康服务模式的革新尝试。

我们未来将持续优化模型精度与问答体验，拓展诊断病种范围，并积极推动系统在真实环境中的部署与落地。

同时，我们计划在未来将项目代码开源，从而降低使用门槛，助力构建“人人可及、人人可用”的心理健康守护网。

精神健康诊断任重道远、困难重重，EASRL-MH不过是众多尝试推进精神健康诊断进程的项目之一，但只要本项目所做出的贡献可以帮助部分患者提前意识到自己的精神健康问题，从而寻求专业诊断和治疗，避免症状的进一步恶化。我们的工作便达到了其意义所在。

精神健康不是“病了才管”的事，而应像量体温一样简单自然。

让EASRL-MH成为这个时代每个人心灵的温度计。

致谢 Acknowledgements

EASRL-MH项目的开发仍将继续。在此，再次感谢任鹏举教授、张文静老师、贺华老师、王柱教授以及Prof. Rothman的指导。同时感谢来自西安交通大学、西北工业大学、西北工业大学附属中学、UC Berkeley、英才计划和Pioneer Academics的各位老师们。

特别感谢中国人民大学附属中学的乔薇因同学为本项目的PPT和展示视频制作做出的重要贡献。

同时感谢我的家人和朋友们在项目的研发过程中对我的情感支持，没有你们，这一切都不可能实现。

“There is no health without mental health.”

—World Health Organization (WHO)