缓解多模态抑郁症检测中的面试官偏见: 一种对抗学习和上下文位置编码的方法

张恩施，克里斯蒂安·波埃拉鲍尔
奈特基金会计算与信息科学学院
美国佛罗里达国际大学，迈阿密，FL 33199
{ezhan004, cpoellab}@fiu.edu

摘要

临床访谈是评估抑郁症的标准方法。最近的方法通过关注面试官提出的特定问题以及手动选择的针对心理健康内容的问答 (QA) 对来提高预测准确性。然而，这些方法往往忽略了更广泛的对话背景，导致泛化能力有限且鲁棒性降低，特别是在结构化程度较低的访谈中，这种访谈在现实世界的临床环境中很常见。在这项工作中，我们开发了一种多模态对话级变压器，通过结合顺序位置嵌入和问题上下文向量来捕捉每次访谈中的对话动态。除了抑郁症预测分支外，我们还构建了一个带有梯度反转层的对抗分类器，以学习对访谈中提出的问题类型保持不变的共享表示。这种方法旨在减少有偏学习，并提高在各种临床访谈场景中抑郁症检测的公平性和泛化能力。在三个基于现实世界访谈的数据集和一个合成数据集上进行的分类和回归实验证明了我们模型的鲁棒性和泛化能力。

1 引言

重度抑郁症，通常称为抑郁症，是一种普遍的心理健康状况，可能会产生严重后果，包括情绪困扰、社交退缩甚至自杀。世界卫生组织 (WHO)[1] 报告称，全球有超过 3 亿人受到抑郁症影响，这对个人、家庭和整个社会都有重大影响。不幸的是，在许多社区，缺乏认识和经济限制等因素导致抑郁症和其他心理健康问题的诊断不足和治疗不足 (世界卫生组织, 2017)。

临床访谈是评估抑郁症状的标准方法 (He 等人, 2022)。每次访谈由面试官和参与者之间的一系列问答 (QA) 对组成。在过去十年中，与单模态方法相比，结合多种数据源 (如访谈期间收集的音频、转录文本和视频) 的多模态方法表现出了更好的性能 (Gong 和 Poellabauer, 2017；Al Hanai 等人, 2018；Yang 等人, 2024；Zhang 等人, 2024)。

最近的多模态研究表明，将面试官的问题作为额外的文本模态纳入可以提高抑郁症检测 (Shen 等人, 2022；Milintsevich 等人, 2023；Chen 等人, 2024；Agarwal 等人, 2024)。然而，报告的高准确率可能源于模型学习面试官的问题模式而不是参与者的回答 (Burdisso 等人, 2024)。例如，一个后续问题，如“你的心理健康状况有改善吗?”可能反映了之前的肯定回答，提供了关于参与者状况的间接线索。一些先前的工作还使用手动选择的问答对来区分抑郁症患者 (Yang 等人, 2017；Agarwal 等人, 2024)，通常忽略了更广泛的对话背景。图 1 展示了来自同一访谈的两个摘录。在图 1a 中，交流揭示了明显的抑郁线索，而图 1 b 的信息较少，这使得前者更容易让模型根据表面信号将其标记为抑郁。

因此，尽管先前工作中呈现的性能指标令人印象深刻，但我们对这些模型在其他基于访谈的语音数据集和现实世界临床情况中的泛化能力和鲁棒性深感担忧，在这些情况下，访谈问题往往更通用，参与者可能会隐瞒他们的真实感受，或者在某些情况下夸大他们的症状 (Pretorius 等人, 2019；Wilson 等人, 2011；Mao 等人, 2023；Zhang 等人, 2025)。重要的是，模型不要依赖这些误导性的捷径进行区分。

在本文中，我们使用一种多模态架构来解决临床访谈中抑郁症检测中面试官引起的偏见和捷径学习问题，该架构将面试官的问题和参与者的回答通过门控融合相结合。我们引入基于对话的上下文位置编码以更好地理解对话轮次，并引入一个轻量级对话变压器来捕捉访谈动态。此外，我们应用一种对抗正则化策略，该策略使用大语言模型 (LLM) 注释的面试官问题函数作为对抗分类器的目标，并且该分类器与

[1] https://www.who.int/news-room/fact-sheets/detail/depression

梯度反转层配对。结果，我们的模型经过训练可以准确预测抑郁症，同时减少其对潜在有偏的面试官提问模式的依赖。代码已发布。[2]

2 相关工作

2.1 单模态抑郁检测

抑郁检测中的单模态方法通常分析单个数据流，最常见的是关注参与者。这些方法包括使用自监督预训练模型 (Zhang 等人，2021 年) 或带有长短期记忆网络 (LSTM) 的传统声学特征 (Du 等人，2023 年) 等技术来处理参与者的语音音频。文本分析探索了转录访谈中的参与者回答，通常使用图卷积网络进行语义理解 (Burdisso 等人，2023 年)，或者通过使用 BERT 定义症状模式 (Nguyen 等人，2022 年) 或使用胶囊网络进行对比学习 (Liu 等人，2024 年) 来利用社交媒体文本。

2.2 多模态抑郁检测

抑郁检测的多模态方法通常整合来自参与者的两个或更多数据流，如语音音频、转录本中的词汇内容和视觉线索，以实现比单模态方法更稳健的评估。研究人员探索了各种技术，包括使用主题建模整合不同特征 (Gong 和 Poellabauer，2017 年)，使用长短期记忆网络对音频 - 文本交互进行序列建模 (Al Hanai 等人，2018 年)，利用自监督基础模型进行丰富表示 (Wu 等人，2023 年)，在融合中应用专家知识 (Yang 等人，2024 年)，使用跨模态注意力 (Iyortsuun 等人，2024 年)，以及将声学地标集成到语言模型中 (Zhang 等人，2024 年)。

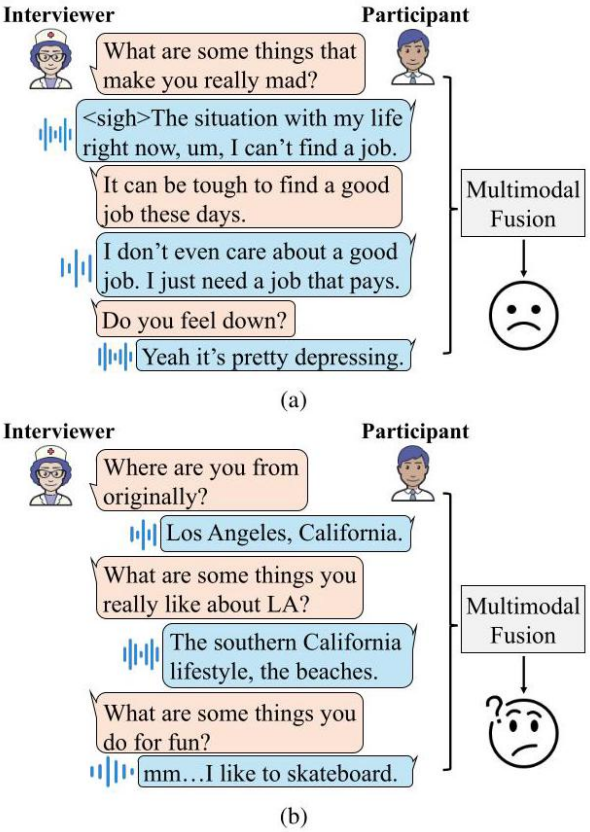

图 1: 来自同一次访谈的两个片段。第一个 (顶部) 片段包含更多用于评估抑郁的判别信息，而第二个 (底部) 则不太清晰。

最近的研究表明，将访谈者的问题作为额外的文本模态纳入可以提高性能 (Niu 等人，2021 年；Dai 等人，2021 年；Shen 等人，2022 年；Milintsevich 等人，2023 年；Agarwal 等人，2024 年；Chen 等人，2024 年；Xue 等人，2024 年)。然而，(Burdisso 等人，2024 年) 发现这样的模型可能利用“捷径”，依赖访谈者的提示而不是参与者的语音或语言，导致在某些数据集上结果虚高，并且在不同访谈风格下的泛化性较差。

我们的多模态框架旨在抵消访谈者引入的偏差。我们通过将序列位置与问题内容合并来使用基于对话的上下文位置编码，以获得包含轮次级上下文的表示。此外，我们使用梯度反转层实现的对抗性访谈者行为正则化，训练模型学习不受潜在偏差提问模式影响的表示。这种方法促进了对抑郁的更稳健且真正以参与者为中心的评估。

3 方法

3.1 问题表述

$\mathcal{D} = \{I_1, \ldots, I_n\}$ 数据集由 n 次临床访谈组成，每次访谈 I 由一系列 k 问答对 $\{(Q_j, A_j)\}_{j=1}^{k}$ 组成，其中 k 在不同访谈中可能会有所不同。每个问题 Q_j 以文本格式呈现，而每个参与者的回答 A_j 是多模态的，包括音频记录 A_j^{audio} 及其相应的转录文本 A_j^{text} 。目标是预测每次完整访谈 I_k 的抑郁状态。这涉及两个预测任务。第一个任务是二元分类，旨在预测抑郁标签 $y_k \in \{0, 1\}$ ，其中 0 表示健康个体，1 表示抑郁个体。第二个任务是回归任务，我们在其中预测一个连续的抑郁标量分数 $y_k \in \mathbb{R}_{\geq 0}$ 。整体框架如图 2 所示。

3.2 特征提取

转录语音。为了从采访者问题和参与者转录文本中提取语义表示,我们使用 XLM-RoBERTa(XLMR)(Conneau 等人，2019)，这是一种基于多语言变换器的语言模型。XLMR 基于 RoBERTa 架构构建，并使用掩码语言建模目标在 100 种不同语言上进行训练 (Conneau 等人，2019)。对于每个输入句子，我们首先对其进行分词和编码，以获得固定维度的向量表示。通过对所有词元嵌入应用平均池化，我们为每个句子导出一个 768 维特征向量的句子级嵌入，问题用 q_i^{text} 表示，参与者的回答用 a_i^{text} 表示。

音频信号。对于参与者的语音回答，我们使用 Wav2Vec2-XLSR-53(XLSR-53)(Conneau 等人，2020) 作为多语言自监督语音编码器来提取音频特征。XLSR-53 是 wav2vec 2.0 large(Baevski 等人，2020) 的一个变体，使用 16kHz 采样语音音频在 53 种语言上进行预训练，包括英语、简体中文和意大利语。XLSR-53 输出维度为 1024 的帧级音频嵌入。我们对时间帧应用平均池化以获得固定维度的表示，用 a_i^{raw} 表示。然后我们应用一个线性投影层将音频嵌入映射到与文本相同的维度空间 (768):

$$a_i^{\text{audio}} = W_{\text{proj}} \cdot a_i^{\text{raw}} + b \tag{1}$$

其中 $a_i^{\text{raw}} \in \mathbb{R}^{1024}$ 是平均池化后的音频嵌入，$W_{\text{proj}} \in \mathbb{R}^{768 \times 1024}$ 是学习到的投影矩阵。

关于选择 XLMR 和 XLSR-53 而不是其他模型的基本原理，以及它们针对每种语言的架构和预训练细节，将在附录 I 中讨论。

[2] https://github.com/coolsoda/e-dep/tree/main

3.3 模态融合

我们采用门控融合机制 (Arevalo 等人，2017) 来组合每个问答对的采访者问题 $\left(q_j^{\text{text}}\right)$、参与者的文本回答 $\left(a_j^{\text{text}}\right)$ 和音频特征 $\left(a_j^{\text{audio}}\right)$。每个模态通过具有 ReLU 激活的两层 MLP 进行编码:

$$h_q = \text{MLP}_q\left(q_j^{\text{text}}\right)$$

$$h_t = \text{MLP}_t\left(a_j^{\text{text}}\right)$$

$$h_a = \text{MLP}_a\left(a_j^{\text{audio}}\right) \tag{2}$$

每个 MLP 映射 $\mathbb{R}^{768} \to \mathbb{R}^{768}$。门控系数使用原始输入的线性投影上的 sigmoid 激活来计算:

$$g_m = \sigma\left(W_m m_j + b_m\right),\ m \in \{q, t, a\} \tag{3}$$

最终融合表示 z_j 是逐元素门控和:

$$z_j = g_q \odot h_q + g_t \odot h_t + g_a \odot h_a \tag{4}$$

3.4 上下文位置编码

上下文位置编码 (CoPE)(Golovneva 等人，2024) 是一种在基于变换器的自然语言处理中使用的方法，它根据上下文调整位置嵌入，而不是基于固定的绝对或相对索引 (Vaswani，2017；Shaw 等人，2018；Raffel 等人，2020)。虽然在诸如计数和选择性复制等任务中有效，但原始的 CoPE 根据内容词元增加位置，不太适合轮次级对话建模。

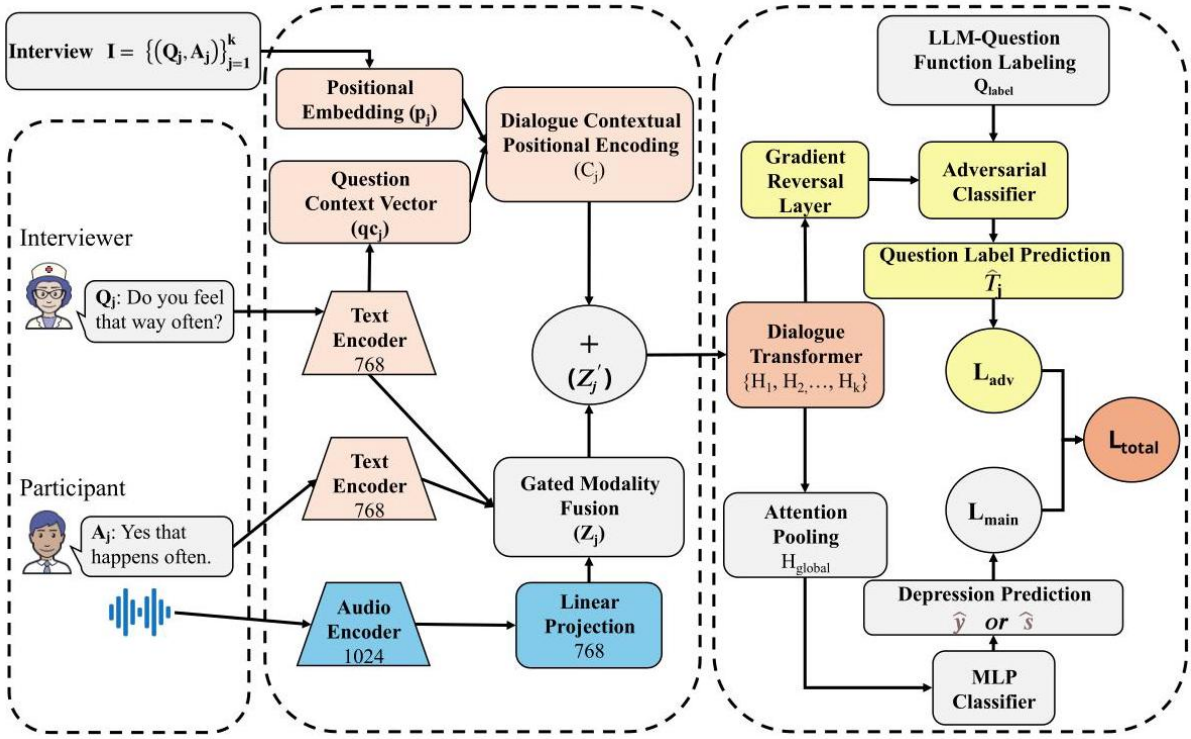

图 2: 所提出的用于抑郁症检测的多模态框架概述。(k) 是当前访谈的长度，(j) 是轮次索引。该模型使用各自的文本和音频编码器处理访谈者问题 (Q_j) 和参与者回答 (文本 A_j^t 、音频 A_j^a)。融合表示 (Z_j) 与基于对话的上下文位置编码 (C_j) 相加，以生成 Z'_j 。然后，一个对话变换器生成上下文轮次嵌入 H_j 。对于抑郁症预测，H_j 序列通过注意力池化 (H_{global}) 进行聚合，并输入到一个多层感知器分类器中。并行地，一个带有梯度反转层的对抗分支使用 H_j 来预测由语言模型注释的访谈者问题功能 (Q_{label}/j) ，从而鼓励对访谈者问题不变的表示。

为了解决这个问题，我们提出了基于对话的 CoPE(D-CoPE)，它是专门为对临床访谈中的问答 (QA) 轮次进行编码而设计的。D-CoPE 整合了轮次位置和访谈者语义。对于长度为 k 的访谈中的每个 QA 对 j ，我们生成一个绝对正弦位置嵌入 $p_j \in \mathbb{R}^{768}$ 。访谈者的问题 q_j^{text} 通过一个两层的多层感知器来提取一个上下文向量:

$$qc_j = \text{MLP}_{\text{CoPE}}\left(q_j^{\text{text}}\right) \tag{5}$$

然后，我们将 p_j 和 qc_j 连接起来，并将结果投影回模型的隐藏空间:

$$c_j = W_{\text{CoPE}}\,[p_j; qc_j] + b_{\text{CoPE}} \tag{6}$$

其中 $W_{\text{CoPE}} \in \mathbb{R}^{768\times 1536}$ 和 $b_{\text{CoPE}} \in \mathbb{R}^{768}$ 。最终的 D-CoPE 向量 c_j 与融合的多模态表示 z_j 按元素相加:

$$z'_j = z_j + c_j \tag{7}$$

3.5 对话级变换器

为了捕捉完整的对话上下文并对访谈中不同 QA 轮次之间的相互依赖关系进行建模，我们将 CoPE 增强的融合表示 $\{z'_1, z'_2, \ldots, z'_k\}$ 序列作为轻量级两层变换器编码器的输入进行处理。序列 $Z' \in \mathbb{R}^{k\times 768}$ 被输入到一个具有 L 层的标准变换器编码器中:

$$Z' \in \mathbb{R}^{k\times 768} \tag{8}$$

编码器的输出是一个上下文表示序列:

$$H = \text{TransformerEncoder}\,(Z') \tag{9}$$

其中 $H \in \mathbb{R}^{k\times 768}$ 。每个向量 H_j 是第 j 个 QA 对的上下文表示。

3.6 抑郁症检测

为了从上下文 QA 嵌入 $\{H_1, \ldots, H_k\}$ 序列中获得整个访谈的固定大小表示，我们应用注意力池化。全局访谈嵌入 H_{global} 被计算为加权和:

$$H_{\text{global}} = \sum_{j=1}^{T_k} \alpha_j H_j \tag{10}$$

其中 α_j 表示第 j 个 QA 对的学习到的注意力权重。这有效地总结了单个访谈中的所有轮次。

H_{global} 然后被输入到一个在隐藏层具有 ReLU 激活的两层多层感知器分类器中。对于二分类任务，最后一层使用 sigmoid 激活来产生一个概率 $\hat{y} \in [0, 1]$ ，该概率表示抑郁症的可能性。该模型使用二元交叉熵损失进行训练:

$$\mathcal{L}_{\text{main}} = -\left[y \log\left(\widehat{y}\right) + (1-y) \log\left(1-\widehat{y}\right)\right] \tag{11}$$

对于回归任务，输出层使用线性激活来产生一个对应于抑郁症严重程度得分的标量预测 $\widehat{s}$。训练目标是均方误差 (MSE) 损失:

$$\mathcal{L}_{\text{main}} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{s}_i - s_i\right)^2 \tag{12}$$

其中 s_i 是第 i 个访谈的真实得分，$\widehat{s}_i$ 是模型的预测。

3.7 提示标签预测

在数据预处理阶段, 对于采访者提出的每个提示 Q_j , 我们使用大语言模型 (LLMs) 将提示分类为七个精心定义的问题功能 (QF) 之一: “开放式”、“转变话题”、“中性信息收集”、“过渡性”、“具体探究”、“支持性”和“其他”。这些 QF 基于问题分类的基础分类法以及临床心理学和心理治疗的功能分类法 (Trzepacz 和 Baker，1993 年；Choi 和 Pak，2004 年；Peräkylä 等人，2008 年；Kallio 等人，2016 年)。附录 H 中讨论了大语言模型遵循的详细提示、说明和原则。

对于每个作为对话级变换器输出的问答级表示 H_j ，我们训练一个对抗分类器。该分类器前面有一个梯度反转层 (GRL)，旨在预测 QF 标签 T_j :

$$\widehat{T}_j = \text{MLP}_{\text{adv}}\left(\text{GRL}\left(H_j\right)\right) \tag{13}$$

MLP_{adv} 是一个具有两层的浅前馈网络。对手试图从 H_j 预测 QF $\left(T_j\right)$ ，并最小化对抗损失 $\mathcal{L}_{\text{adv}}$ ，因此 $\mathcal{L}_{\text{adv}}$ 的梯度流回对手，使其更擅长从 H_j 预测 QF。

GRL 位于主要抑郁预测模型和对手之间。在反向传播过程中，当来自 $\mathcal{L}_{\text{adv}}$ 的梯度到达 GRL 时，它会翻转梯度的符号。因此，主要抑郁预测模型会收到两组梯度信号来指导 H_j 的形成。从损失函数 $\mathcal{L}_{\text{main}}$ 中，模型学习改进 H_j 以预测抑郁。相比之下，损失函数 $\mathcal{L}_{\text{adv}}$ 旨在使 H_j 在预测 QF 时效果降低。这种双重方法要求主要模型识别与 QF 信息无关的抑郁预测特征。

3.8 训练目标

我们模型的总体训练目标旨在同时实现两个目标: 准确预测主要结果 (例如，抑郁状态) 并减少模型对潜在有偏差的采访者问题功能 (QF) 的依赖。这是通过将主要预测损失 ($\mathcal{L}_{\text{main}}$) 与具有超参数 λ 的对抗正则化损失 ($\mathcal{L}_{\text{adv}}$) 相结合来实现的:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} - \lambda \cdot \mathcal{L}_{\text{adv}} \tag{14}$$

超参数 λ 是一个非负标量，它控制着在最小化主要任务的预测误差和最小化学到的表示 H_j 中关于 QF 的信息之间的权衡。

4 实验设置

4.1 数据增强和预处理

在本研究中，我们使用了三个真实世界数据集和一个合成数据集，并且所有数据集都经过了完全匿名化处理。痛苦分析访谈语料库/奥兹国巫师 (DAIC-WOZ) 数据集 [3] (格拉奇等人，2014 年) 是基于语音的

抑郁症研究中使用最广泛的资源之一。这个英语数据集包含从 189 名参与者收集的 189 次访谈。它具有 16kHz 音频记录、转录文本和手动提取的视觉数据。根据 8 项患者健康问卷 (PHQ - 8)(克伦克等人，2009 年) 为每个参与者分配一个分数。在这 189 次访谈中，57 名参与者被归类为抑郁症患者，而 132 名被归类为健康对照。情感音频 - 文本抑郁症 (EATD) 语料库 [4] (沈等人，2022 年) 是一个中文数据集，包含对 162 名参与者的访谈，每个访谈都使用自评抑郁量表 (SDS)(宗，1965 年) 进行标注。在这个数据集中，30 名参与者被确定为抑郁症患者，而 132 名被归类为健康对照。它还包括 16kHz 的音频记录及其相应的转录文本。安卓语料库 [5] (陶等人，2023 年) 由对 116 名参与者的 116 次访谈组成，访谈用意大利语进行，并根据《精神疾病诊断与统计手册》(DSM - 5)(协会等人，2013 年) 进行标注。在这个数据集中，62 名参与者被确定为抑郁症患者，54 名被确定为健康对照。音频记录最初为 44.1 kHz，并已重新采样为 16kHz。三个数据集的详细信息在附录 C 中讨论。

临床数据集常常呈现出不平衡的类别分布，这会显著影响模型性能。对于 DAIC - WOZ 数据集，我们在训练集中应用随机过采样，为少数 (抑郁症) 类别创建重复样本，从而在抑郁症患者数量和健康对照之间实现平衡。对于 EATD 语料库，我们遵循原始作者提出的过采样策略，重新排列每次访谈中问答对的顺序，生成额外的数据点以平衡类别分布 (沈等人，2022 年)。在 DAIC - WOZ 和 EATD 数据集中，随机过采样严格应用于训练集，以确保现实的验证并防止任何数据泄漏。安卓语料库最初是以平衡的方式构建的，所以没有应用过采样。

鉴于数据集规模有限，我们使用大语言模型 (GPT - 4o)(赫斯特等人，2024 年) 按照与先前研究 (陈等人，2024 年) 类似的方法合成额外的文本数据用于训练。由于 DAIC - WOZ 数据集的规模及其在该领域用于研究和基准测试挑战的普遍使用，合成数据基于该数据集的每个训练集 (称为 DAIC - 合成)。对于训练集中的每次访谈，访谈者的问题保持不变，并指示大语言模型以文本形式生成参与者回答的替代内容。这些回答经过重新措辞以保持原始内容和词汇的使用。我们通过随机帧交换基于原始音频创建参与者音频的新样本。由于 DAIC - WOZ 数据集中没有原始视频数据，在实现利用视觉模态的基线模型时，我们为每次访谈复制了现有的低级视觉特征。提供给大语言模型的详细说明和提示可在附录 B 中找到。

执行细节

所提出的框架使用 PyTorch 实现。该模型在谷歌 Colab Pro 上使用 NVIDIA A100 GPU 进行训练和评估，该 GPU 的驱动版本为 550.90.12，CUDA 版本为 12.4，系统内存为 83.5 GB，GPU 内存为 40 GB。该模型以 8 的批量大小训练 20 个轮次，使用 AdamW 优化器 (金马，2014 年)，初始学习率为 1×10^{-3}。

每个模块的大多数超参数在第 3 节中讨论。每个模块所有超参数及其探索范围的综合列表可在附录 F 中找到。

评估

为保持一致性并便于与每个数据集中的基线进行公平比较，我们使用五折分层交叉验证进行评估。对于每个数据集，所有数据点使用分层分割分为五折，以确保每一折与原始数据集保持相同的类别分布 (即抑郁和健康参与者的比例)。在每次训练轮次中，一折用作测试集，其余四折用于训练。此过程重复五次，每一折都用作一次测试集。对于每个训练集，我们应用第 4.1 节中讨论的过采样策略来平衡类别分布。最终性能报告为五次独立训练-测试评估的平均值。

我们使用四个指标评估模型，以评估其分类和回归性能。对于确定参与者是否属于阳性类别 (抑郁) 的二元分类任务，我们报告平衡精确率和召回率的 F1 分数，以及评估模型在不同决策阈值下区分类别的能力的 AUC-ROC 分数。对于预测 PHQ-8 分数的回归任务，我们使用均方根误差 (RMSE) 来衡量预测误

[3] https://dcapswoz.ict.usc.edu/

差的大小，并使用平均绝对误差 (MAE) 来提供预测分数与实际分数之间平均偏差的补充度量，该度量考虑了值的规模。附录 D 中提供了每个指标的详细计算。

模型	模态	DAIC-WOZ		EATD		安卓机器人		DAIC 合成数据	
		F1	均方根误差	F1	均方根误差	F1	均方根误差	F1	均方根误差
GCN-PE(布尔迪索等人，2024 年)	I	0.84	4.51	0.55	5.94	0.59	-	0.84	4.45
WU(吴等人，2023 年)	A + T	0.80	4.36	0.67	4.21	0.71	-	0.81	4.29
MMPF(杨等人，2024 年)	A + T	0.76	5.11	0.66	4.46	0.70	-	0.78	5.02
ACMA(约尔楚恩等人，2024 年)	A + T	0.69	5.15	0.61	5.12	0.68	-	0.72	4.80
大语言模型 (张等人，2024 年)	A + T	0.76	5.04	0.64	4.68	0.69	-	0.79	4.61
CAMFM(薛等人，2024 年)	A + T	0.80	4.71	0.70	4.10	0.72	-	0.81	4.51
DAI(戴等人，2021 年)	I + A + T + V	0.81	4.67	0.67	4.33	0.71	-	0.80	4.65
HCAG(牛等人，2021 年)	I + A + T	0.80	4.79	0.65	5.54	0.69	-	0.81	4.43
SHEN(沈等人，2022 年)	I + A + T + V	0.77	5.25	0.68	4.59	0.70	-	0.77	5.14
MILI(米林采维奇等人，2023 年)	I + T	0.75	5.11	0.62	5.97	0.68	-	0.76	4.96
SEGA(陈等人，2024 年)	I + A + T + V	0.71	5.04	0.70	4.93	0.70	-	0.74	4.90
GCN(布尔迪索等人，2023 年)	I + T	0.79	4.95	0.61	5.52	0.69	-	0.80	4.87
AGAR(阿加瓦尔等人，2024 年)	I + T	0.72	5.96	0.58	5.53	0.67	-	0.72	5.91
对话变换器 (我们的)	I + A + T	0.82	3.86	0.72	4.04	0.73	-	0.83	3.84

表 1: 在每个数据集上单独进行训练和测试的结果。A, T 和 V 分别指参与者音频、转录文本和视觉特征；I 表示采访者提示 (文本)。最佳结果用粗体表示，次佳结果用下划线表示，“-”表示无可用数据。F1 越高且均方根误差 (RMSE) 越低表明性能越好。

4.4 基线和消融实验

我们从三个角度将我们的方法与基线进行比较。首先，我们纳入了明确纳入采访者问题的多模态模型，因为这些问题为解释参与者的回答提供了重要背景。其次，我们考虑了排除采访者输入的强大多模态基线。第三，据我们所知，我们纳入了一项仅使用采访者问题在 DAIC-WOZ 基准上取得了当前最优性能的研究。所有基线都遵循第 4.3 节中描述的相同训练和评估设置，每个基线模型在附录 G 中详细讨论。

我们的消融实验分为两类。第一类重点是在我们用于抑郁症检测任务的模型中消融关键模块，第二类评估每个单独数据模态的贡献。

5 结果与讨论

5.1 整体性能

如第 4.4 节所述，我们将我们的方法与三个基线组进行了比较，并将结果列于表 1 中。我们的观察表明，GCN-PE，一种专门设计用于利用采访者问题线索的方法，在 DAIC-WOZ 和 DAIC-Synthetic 数据集中取得了最高的 F1 分数。然而，它在 Androids 数据集上的性能显著下降，在 EATD 数据集上下降得更多。这两个数据集包含的采访较短，采访者的提示更通用且结构较少，这表明 GCN-PE 在不同的采访环境中缺乏通用性。

我们发现，纳入采访者问题 (模态 I) 的多模态方法在 DAIC-WOZ 和 DAIC-Synthetic 数据集中通常表现良好。然而，当应用于 EATD 和 Androids 时，它们的性能有时会显著下降。这表明 DAIC-WOZ 中的采访者问题比其他数据集中的问题提供了更多信息且更一致。因此，依赖这种模态的模型在采访者行为更

[4] https://github.com/speechandlanguageprocessing/ ICASSP2022-Depression

[5] https://github.com/androidscorpus/data

通用或更多样化的场景中往往表现不佳。这一趋势也反映在我们提出的方法中，与 DAIC-WOZ 相比，其在 EATD 和 Androids 上的性能明显下降。

对于不使用采访者提示的方法，我们观察到在 DAIC-Synthetic 数据集上性能略好。由于训练数据更平衡且可用样本数量增加，这种改进是预期的。然而，对于使用采访者提示 (I) 和视觉线索 (V) 的模型，在 DAIC-Synthetic 中性能没有显著提高。这表明合成的 (即复制的) 视觉特征带来的益处最小，甚至可能阻碍性能。因此，需要更有效的技术来合成低级视觉特征。

型号变体	模态	DAIC-WOZ		EATD		安卓机器人		DAIC 合成数据	
		F1	均方根误差	F1	均方根误差	F1	均方根误差	F1	均方根误差
D-CoPE + QF + GRL(完整模型)	I + A + T	0.82	3.86	0.72	4.04	0.73	-	0.83	3.84
D-CoPE + QF	I + A + T	0.89	3.81	0.71	4.12	0.70	-	0.90	3.82
D-CoPE + GRL	I + A + T	0.83	3.86	0.72	4.01	0.73	-	0.84	3.74
QF + GRL	I + A + T	0.65	5.41	0.54	4.90	0.58	-	0.63	5.37
基础模型	I + A + T	0.65	5.57	0.53	4.95	0.57	-	0.67	5.40
完整模型	I + A	0.70	4.52	0.61	5.38	0.63	-	0.74	4.31
完整模型	I + T	0.77	4.34	0.65	4.99	0.66	-	0.79	4.19
完整模型	A + T	0.70	4.50	0.70	4.17	0.65	-	0.72	4.56

表 2: 四个数据集上抑郁症检测的消融研究结果。D-CoPE 指我们提出的基于对话的上下文位置编码。QF 表示使用面试官问题功能标签作为对抗分类器的目标，GRL 指对抗训练中使用的梯度反转层。“基础模型”缺少所有这三个组件，但使用相同的底层多模态架构。I, A 和 T 分别表示文本格式的面试官提示、参与者的音频回复、参与者的文本回复。

关于安卓语料库，由于它使用 DSM-5 诊断抑郁症，它只提供二元标签 (抑郁或非抑郁)(如 4.1 节所述)。因此，由于没有连续的目标值，无法进行回归分析。附录 E 包含模型在 AUC-ROC 和 MAE 指标上的其他实验结果。为了进一步评估模型在不同面试官风格、领域和语言变体上的性能，我们进行了全面的跨数据集验证。结果报告在附录 E 的表 4 中。

5.2 消融研究

消融结果总结在表 2 中。D-CoPE 起着关键作用: 去除它会导致 F1 和 RMSE 大幅下降，特别是在 DAIC-WOZ 和安卓等结构化面试数据集上。这证实了 D-CoPE 有效地为对话级变换器编码了序列和问题级语义。

去除 GRL 使模型能够利用与 QF 的相关性，在 DAIC-WOZ 和 DAIC-合成数据集上获得了最高的 F1 和近乎最佳的 RMSE 分数，甚至超过了完整模型。这表明 QF 在这些数据集中是强大的捷径特征。虽然去除 GRL 提高了在共享这些偏差的测试集上的性能，但我们的完整模型通过包含 GRL，有意减少了对这种捷径的依赖。尽管这在原始性能上有一定代价，但它促进了泛化和公平性，而这在测试指标中并未完全体现。

有趣的是，省略针对 QF 标签的对抗损失的变体 $CoPE + GRL\,(NoQF)$，其性能与完整模型相似或略好。这意味着对抗训练引入了有用的正则化，但可能会略微降低分布内准确率以利于去偏。

模态消融进一步证明了每个输入流的贡献。去除参与者文本或音频会降低所有数据集的性能，文本通常更关键。去除面试官问题在 DAIC-WOZ 和安卓数据集上显著降低了性能，突出了它们在将参与者回复上下文化和支持 D-CoPE 方面的作用。在使用通用提示的 EATD 数据集上，去除面试官输入的影响不太明显。

6 结论

在这项工作中，我们开发了一个多模态框架来增强从临床语音访谈数据中检测抑郁症的能力。主要目标是确保模型从参与者的回复中学习到有意义的抑郁症表示，而不是依赖于面试官问题的表面线索或过度拟合到特定的、手动探究的问答对，这可能导致误导性的性能提升。结果突出了明确建模和解决临床数据收集过程中可能出现的潜在偏差的重要性。我们希望我们的工作有助于构建一个更可靠的解决方案，以开发用于心理健康评估的公平且可泛化的人工智能系统。

局限性

我们从两个角度讨论我们工作的局限性: 一个与使用的数据有关，另一个与模型架构有关。

- 数据。从数据角度来看有两个主要局限性。第一个是可用数据集的大小。有标签的临床语音数据集通常很小且受隐私法规限制，限制了公众访问。这个挑战不仅限于抑郁症研究，还扩展到其他精神和神经退行性疾病。为了解决这个问题，我们应用了过采样并使用合成生成的数据来提高泛化能力。然而，未来的工作将受益于更大、更多样化且公开可用的数据集，这些数据集可以在现实世界环境中收集或通过先进模型生成。

第二个局限性是模态的范围。由于可用性限制，我们的研究集中在文本和音频上。EATD 和安卓数据集完全缺乏视频数据，虽然 DAIC-WOZ 包含一些视觉特征，但原始视频无法访问。因此，实现与我们的文本/音频设计相当的视觉管道是不可行的。此外，我们的工作集中在减轻音频文本数据中的面试官偏差，使其他模态超出了我们的范围。然而，我们承认整合视觉 (如面部表情、医学成像) 和生理信号 (如心率、皮肤电导) 以丰富用于抑郁症检测的多模态模型的价值。随着更全面的数据集出现，这仍然是一个有前途的方向。

- 模型。我们的框架是针对二元临床访谈量身定制的，利用基于对话的上下文位置编码和针对面试官问题功能 (IQF) 的对抗正则化。虽然该模型对结构化访谈有效，但可能无法推广到独白或自我报告等非交互式情境，在这些情境中关键组件会失去功能。使我们的设计适应此类格式是一个单独的研究挑战。

IQF 定义是我们减轻偏差策略的核心。我们将这些类别基于临床访谈和对话行为分类法，使用大语言模型在上下文中标记问题。虽然这种方法能够实现可扩展的标注，但可能无法完全捕捉真实面试官行为的细微差别。大语言模型的输出还取决于基础模型和提示策略，这可能会限制精度。未来的工作可以使用人工标注员来涵盖更广泛的面试官行为标签，并利用可用资源帮助验证大语言模型生成标签的准确性。

参考文献

Navneet Agarwal、Gaël Dias 和 Sonia Dollfus。2024 年。分析话语结构对改善心理健康评估的相关性。见第 9 届计算语言学与临床心理学研讨会 (CLPsych 2024) 论文集，第 127 - 132 页。

Tuka Al Hanai、Mohammad M Ghassemi 和 James R Glass。2018 年。通过访谈的音频/文本序列建模检测抑郁症。见国际语音通信协会年会 (Interspeech)，第 1716 - 1720 页。

Relja Arandjelovic、Petr Gronat、Akihiko Torii、Tomas Pajdla 和 Josef Sivic。2016 年。Netvlad: 用于弱监督地点识别的卷积神经网络架构。见电气和电子工程师协会计算机视觉与模式识别会议论文集，第 5297 - 5307 页。

John Arevalo、Thamar Solorio、Manuel Montes - y Gómez 和 Fabio A González。2017 年。用于信息融合的门控多模态单元。arXiv 预印本 arXiv:1702.01992。

美国精神病学协会等。2013 年。精神疾病诊断与统计手册: $DSM-5$。美国精神病学协会。

Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。2020 年。wav2vec 2.0: 语音表示自监督学习框架。神经信息处理系统进展，33:12449 - 12460。

Iz Beltagy、Matthew E Peters 和 Arman Cohan。2020 年。Longformer: 长文档变换器。arXiv 预印本 arXiv:2004.05150。

Sergio Burdisso、Ernesto Reyes - Ramírez、Esaú Villatoro - Tello、Fernando Sánchez - Vega、Pastor López - Monroy 和 Petr Motlicek。2024 年。Daic - woz: 关于在临床访谈自动抑郁症检测中使用治疗师提示的有效性。arXiv 预印本 arXiv:2404.14463。

Sergio Burdisso、Esaú Villatoro - Tello、Srikanth Madik - eri 和 Petr Motlicek。2023 年。用于转录临床访谈中抑郁症检测的节点加权图卷积网络。arXiv 预印本 arXiv:2307.00920。

陈三元、王承义、陈正阳、吴宇、刘树杰、陈卓、李锦玉、神田直之、吉冈卓也、肖雄等。2022 年。Wavlm: 用于全栈语音处理的大规模自监督预训练。IEEE 信号处理杂志精选主题，16(6):1505 - 1518。

陈庄、邓家文、周金峰、吴锦岑、钱铁云、黄敏丽。2024 年。基于大语言模型赋能的结构元素图的临床访谈抑郁症检测。见北美计算语言学协会 2024 年会议: 人类语言技术 (第 1 卷: 长论文) 论文集，第 8174 - 8187 页。

Bernard CK Choi 和 Anita WP Pak。2004 年。问卷偏差目录。预防慢性病，2(1):A13。

Alexis Conneau、Alexei Baevski、Ronan Collobert、Abdelrahman Mohamed 和 Michael Auli。2020 年。用于语音识别的无监督跨语言表示学习。arXiv 预印本 arXiv:2006.13979。

亚历克西斯·科诺、卡尔蒂凯·坎德瓦尔、纳曼·戈亚尔、维什拉夫·乔杜里、纪尧姆·温泽克、弗朗西斯科·古兹曼、爱德华·格雷夫、迈尔·奥特、卢克·泽特尔-莫耶尔和韦塞林·斯托亚诺夫。2019 年。大规模无监督跨语言表征学习。arXiv 预印本 arXiv:1911.02116。

马克·G·科尔和詹姆斯·艾伦。1997 年。使用 damsl 注释方案对对话进行编码。在 AAAI 关于人类与机器中的交际行为的秋季研讨会上，第 56 卷，第 28 - 35 页。马萨诸塞州波士顿。

戴志军、周恒、巴清芳、周扬、王立峰和李国臣。2021 年。使用一种结合上下文感知分析的新型特征选择算法改进抑郁症预测。《情感障碍杂志》，295:1040 - 1048。

雅各布·德夫林、张明伟、肯顿·李和克里斯蒂娜·图托纳娃。2018 年。Bert: 用于语言理解的深度双向变换器的预训练。arXiv 预印本 arXiv:1810.04805。

杜明浩、刘爽、王涛、张文泉、柯宇峰、陈龙和董明。2023 年。使用融合语音产生和感知特征的语音链模型进行抑郁症识别。《情感障碍杂志》，323:299 - 308。

弗洛里安·艾本、克劳斯·R·舍雷尔、比约恩·W·舒勒、约翰·桑德伯格、伊丽莎白·安德烈、卡洛斯·布索、劳伦斯·Y·德维勒、朱利安·埃普斯、佩特里·劳卡、什里坎特·S·纳拉亚南等人。2015 年。用于语音研究和情感计算的日内瓦简约声学参数集 (gemaps)。《IEEE 情感计算汇刊》，7(2):190 - 202。

弗洛里安·艾本、马丁·沃尔默和比约恩·舒勒。2010 年。开源多功能快速音频特征提取器 Opensmile。发表于第 18 届 ACM 国际多媒体会议论文集，第 1459 - 1462 页。

奥尔加·戈洛夫涅娃、王天禄、杰森·韦斯顿和赛音巴亚尔·苏赫巴特尔。2024 年。上下文位置编码: 学习计算重要的内容。arXiv 预印本 arXiv:2405.18719。

袁公和克里斯蒂安·波埃拉鲍尔。2017 年。基于主题建模的多模态抑郁症检测。载于《第 7 届音频/视觉情感挑战年度研讨会论文集》，第 69 - 76 页。

乔纳森·格拉奇、罗恩·阿特斯坦、盖尔·M·卢卡斯、乔塔·斯特拉托、斯特凡·舍雷尔、安吉拉·纳扎里安、雷切尔·伍德、吉尔·博贝格、大卫·德沃尔特、斯泰西·马塞拉等人。2014 年。《人类与计算机访谈的痛苦分析访谈语料库》。载于 $LREC$ ，第 3123 - 3128 页。雷克雅未克。

何朗，牛明月，普拉亚格·蒂瓦里，佩卡·马尔蒂宁，苏锐，蒋杰伟，郭晨光，王宏宇，丁松涛，王中民等。2022 年。基于视听线索的抑郁症识别深度学习综述。《信息融合》，80:56 - 86。

许伟宁、本杰明·博尔特、蔡耀宏、库沙尔·拉霍蒂亚、鲁斯兰·萨拉胡丁诺夫和阿卜杜勒-拉赫曼·穆罕默德。2021 年。休伯特: 通过隐藏单元的掩码预测进行自监督语音表示学习。《IEEE/ACM 音频、语音和语言处理汇刊》，29:3451 - 3460。

亚伦·赫斯特、亚当·勒勒、亚当·P·古彻、亚当·佩雷尔曼、阿迪蒂亚·拉梅什、艾丹·克拉克、AJ·奥斯特罗、阿基拉·韦利欣达、艾伦·海斯、亚历克·拉德福德等人。2024 年。Gpt - 4o 系统卡。arXiv 预印本 arXiv:2410.21276。

恩古米米·卡伦·约尔图恩、苏-亨·金、亨-郑·杨、承元·金和闵·约翰。2024 年。基于音频和文本特征的用于抑郁症检测的加法跨模态注意力网络 (acma)。《IEEE 接入》。

汉娜·卡利奥、安娜-迈娅·皮蒂拉、马丁·约翰逊和玛丽·坎加斯涅米。2016 年。系统方法学综述: 为定性半结构化访谈指南制定框架。《高级护理杂志》，72(12):2954 - 2965。

迪德里克·P·金马。2014 年。《亚当: 一种随机优化方法》。arXiv 预印本 arXiv:1412.6980。

库尔特·克伦克、塔拉·W·斯特林、罗伯特·L·斯皮策、珍妮特·BW·威廉姆斯、乔伊斯·T·贝里和阿里·H·莫克-达德。2009 年。《PHQ-8 作为一般人群当前抑郁程度的测量工具》。《情感障碍杂志》，114(1 - 3):163 - 173。

刘汉、李长亚、张晓彤、张峰、王巍、马凤龙、陈宏阳、于洪、张显超。2024 年。《基于对比学习的胶囊网络进行抑郁检测》。收录于《AAAI 人工智能会议论文集》，第 38 卷，第 22231 - 22239 页。

刘音涵、迈尔斯·奥特、纳曼·戈亚尔、杜静飞、曼达尔·乔希、陈丹琦、奥默·利维、迈克·刘易斯、卢克·泽特勒莫耶、维塞林·斯托扬诺夫。2019 年。《RoBERTa: 一种经过稳健优化的 BERT 预训练方法》。arXiv 预印本 arXiv:1907.11692。

毛凯宁、吴宇琦、陈杰。2023 年。《自动临床抑郁诊断的系统综述》。《npj 心理健康研究》，2(1):20。

基里尔·米林采维奇、凯丽特·西尔茨、盖尔·迪亚斯。2023 年。《通过症状预测实现基于文本的抑郁自动估计》。《大脑信息学》，10 (1) : 4 。

威廉·R·米勒、斯蒂芬·罗尔尼克。2012 年。《动机性访谈: 帮助人们改变》。吉尔福德出版社。

阮通、安德鲁·耶茨、阿亚·齐里克利、巴特·德梅特、阿尔曼·科汉。2022 年。《利用临床问卷提高抑郁检测的泛化能力》。arXiv 预印本 arXiv:2204.10432。

牛萌、陈凯、陈庆才、杨陆峰。2021 年。《HCAG: 一种用于抑郁检测的分层上下文感知图注意力模型》。收录于《ICASSP 2021 - 2021 年 IEEE 国际声学、语音和信号处理会议 (ICASSP)》，第 4235 - 4239 页。IEEE。

安西·佩拉基拉、查尔斯·安塔基、桑娜·韦赫维莱宁、伊万·柳德。2008 年。《分析实践中的心理治疗》。收录于《对话分析与心理治疗》，第 5 - 25 页。剑桥大学出版社。

丹尼尔·波维、张晓辉、桑吉夫·库丹普尔。2014 年。《基于自然梯度和参数平均的深度神经网络并行训练》。arXiv 预印本 arXiv:1410.7455。

克劳德特·普雷托里乌斯、德里克·钱伯斯、大卫·科伊尔。2019 年。《年轻人的在线求助与心理健康问题: 系统叙事综述》。《医学互联网研究杂志》，21(11):e13873。

詹姆斯·O·普罗查斯卡、韦恩·F·维利泽。1997 年。《健康行为改变的跨理论模型》。《美国健康促进杂志》，12(1):38 - 48。

亚历克·拉德福德、郑宇锡、徐涛、格雷格·布罗克曼、克里斯汀·麦克利维、伊利亚·苏茨克维。2023 年。《通过大规模弱监督实现鲁棒语音识别》。收录于《国际机器学习会议》，第 28492 - 28518 页。PMLR。

科林·拉菲尔、诺姆·沙泽尔、亚当·罗伯茨、凯瑟琳·李、沙兰·纳朗、迈克尔·马泰纳、周彦琦、李威、彼得·J·刘。2020 年。《用统一的文本到文本变换器探索迁移学习的极限》。《机器学习研究杂志》，21(140):1 - 67。

彼得·肖、雅各布·乌兹科雷特、阿希什·瓦斯瓦尼。2018 年。《具有相对位置表示的自注意力》。arXiv 预印本 arXiv:1803.02155。

应申、杨慧宇和林琳。2022 年。自动抑郁症检测: 一个情感音频-文本语料库和基于 gru/bilstm 的模型。发表于《ICASSP 2022 - 2022 年 IEEE 国际声学、语音和信号处理会议 (ICASSP)》，第 6247 - 6251 页。IEEE。

约翰·萨默斯 - 弗拉纳根和丽塔·萨默斯 - 弗拉纳根。2012 年。临床访谈:2012 - 2013 年更新版。约翰·威利父子公司。

陶福祥、安娜·埃斯波西托和亚历山德罗·温恰雷利。2023 年。安卓语料库: 一个新的公开可用的基于语音的抑郁症检测基准。《抑郁症》，47:11 - 9。

宝拉·T·特泽帕克和罗伯特·W·贝克。1993 年。《精神科精神状态检查》。牛津大学出版社。

A·瓦斯瓦尼。2017 年。《注意力是你所需要的一切》。《神经信息处理系统进展》。

阿努拉德哈·韦利维塔和珀尔·普。2020 年。人类社交对话中移情反应意图的分类法。arXiv 预印本 arXiv:2012.04080。

科拉莉·J·威尔逊、黛布拉·J·里克伍德、约翰·A·布什内尔、彼得·卡普蒂和苏珊·J·托马斯。2011 年。自主性需求和对从非正式来源寻求帮助的偏好对新兴成年人获取常见精神障碍和自杀念头心理健康服务意图的影响。《心理健康进展》，10(1):29 - 38。

世界卫生组织。2017 年。《抑郁症及其他常见精神障碍: 全球健康估计》。

吴文、张超和菲利普·C·伍德兰。2023 年。基于语音的抑郁症检测中的自监督表示。发表于《ICASSP 2023 - 2023 年 IEEE 国际声学、语音和信号处理会议 (ICASSP)》，第 1 - 5 页。IEEE。

薛俊奇、秦瑞涵、周新旭、刘洪海、张敏和张治国。2024 年。融合音频的多级特征和基于文本的上下文句子嵌入用于基于访谈的抑郁症检测。发表于《ICASSP 2024 - 2024 年 IEEE 国际声学、语音和信号处理会议 (ICASSP)》，第 6790 - 6794 页。IEEE。

杨彪、曹苗苗、朱贤林、王苏红、杨长春、倪荣荣和刘晓峰。2024 年。Mmpf: 用于自动抑郁症检测的多模态净化融合。《IEEE 计算社会系统汇刊》。

杨乐、蒋冬梅、夏晓涵、裴尔成、梅西亚·塞德里克·奥韦内克和希谢姆·萨利。2017 年。使用深度学习模型进行抑郁症的多模态测量。发表于《第 7 届年度音频/视觉情感挑战研讨会论文集》，第 53 - 59 页。

张恩施、拉斐尔·M·特鲁希略和克里斯蒂安·波埃拉鲍尔。2025 年。参与者参与度和数据质量: 从心理健康众包感知研究中吸取的教训。《ACM 人机交互汇刊》，9(2):1 - 30。

张平月、吴梦月、海因里希·丁克尔和俞凯。2021 年。Depa: 用于抑郁症检测的自监督音频嵌入。发表于《第 29 届 ACM 国际多媒体会议论文集》，第 135 - 143 页。

张翔宇、刘鹤心、许开帅、张启全、刘黛娇、比娜·艾哈迈德和朱利安·埃普斯。2024 年。当大语言模型遇到声学地标: 一种将语音集成到用于抑郁症检测的大语言模型中的有效方法。arXiv 预印本 arXiv:2402.13276。

威廉·W·K·宗。1965 年。《自评抑郁量表》。《普通精神病学档案》，12(1):63 - 70。

附录

A 概述

附录的结构如下。B 部分详细介绍了我们如何使用大语言模型 (LLM) 来合成额外的数据 (DAIC-合成数据) 用于训练和评估。C 部分提供了关于本研究中使用的每个真实世界数据集的更多信息。D 部分列出了本研究中使用的主要评估指标。E 部分包括在另外两个指标上进行的实验，以及在四个数据集上的跨数据集验证。F 部分概述了一般训练和框架内每个模块的详细超参数。G 部分描述了基线模型的实现细节。H 部分讨论了我们如何使用大语言模型来标注问题功能，I 部分介绍了本研究中用于特征提取的基础模型背后的原理。

B 合成数据

在图 3 中，我们展示了用于通过 GPT-4 API 调用指导大语言模型根据 DAIC-WOZ 数据集中的访谈合成参与者回复的提示模板。每次访谈由多个问答对组成。对于每个问答对，我们促使大语言模型生成三个替代文本回复。

对于模型训练和评估 (如 4.1 节所述)，我们使用分层分割的 5 折交叉验证策略，以确保每一折都保留原始的类别分布。在每一折中，由数据的 80% 组成的训练集通过合成额外的样本进行扩充。这种扩充不

仅增加了训练集的大小，还改善了类别平衡，最终提高了模型的泛化能力。扩充后，每个训练集的大小从 152 个样本增加到 396 个样本。

C 数据收集和预处理的细节

痛苦分析访谈语料库/奥兹巫师 (DAIC-WOZ) 数据集 (Gratch 等人，2014 年) 是抑郁症研究中使用最广泛的数据集之一。这个英语数据集包括对 189 名参与者进行的 189 次访谈，并收集了 16kHz 音频记录、转录文本和视觉数据。它是从大洛杉矶地区的两组参与者中收集的。一组由美国武装部队的退伍军人组成，另一组包括普通公众。每次访谈都涉及一名参与者和一个名为埃莉的人工控制代理。每个参与者都被标记有一个 8 项患者健康问卷 (PHQ-8)(Kroenke 等人，2009 年) 得分，范围从 0 到 24。得分 10 分或以上的参与者被归类为抑郁症患者，得分低于 10 分的参与者被归类为健康对照。在 189 次访谈中，57 名参与者被标记为抑郁症患者，132 名被归类为健康对照。

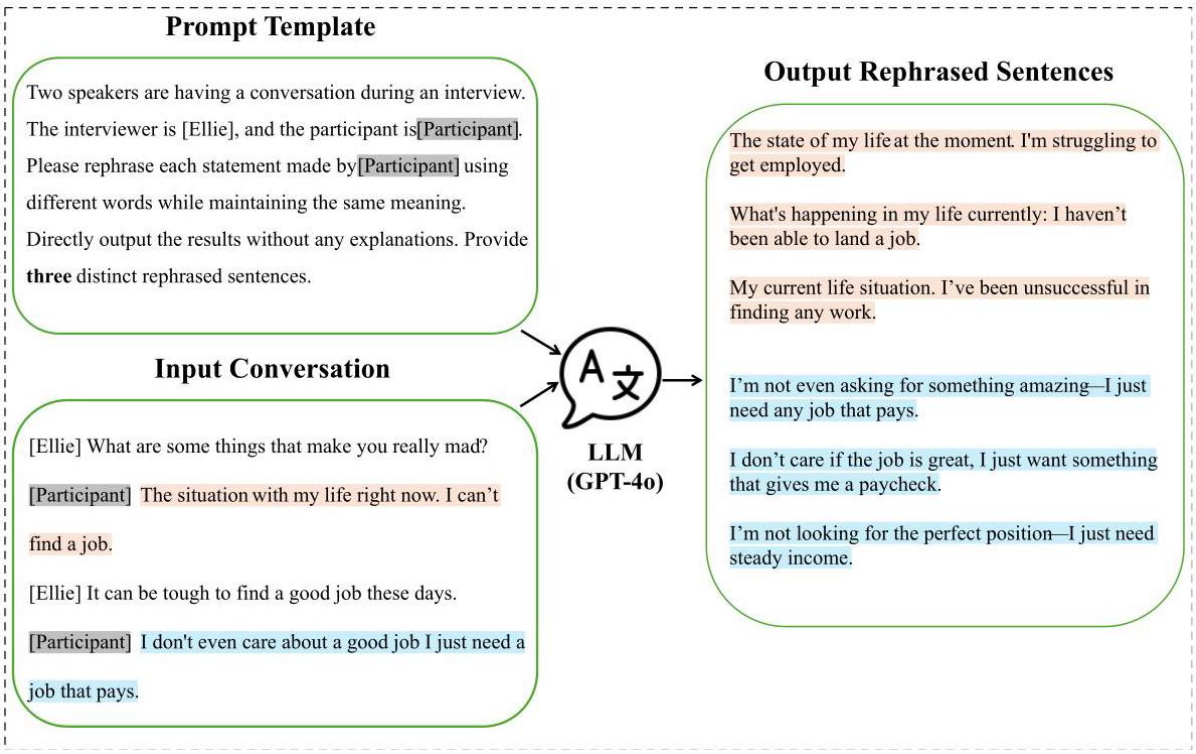

图 3: 使用大语言模型通过重新表述每个问答对中参与者的回复来生成合成数据。

情感音频-文本抑郁症 (EATD) 语料库 (Shen 等人，2022 年) 是一个中文数据集，由对 162 名学生志愿者的访谈组成。其开发的一个动机是通过研究参与者对随机的、设计不太具体的心理健康评估问题的回答来检测抑郁症。访谈以 16kHz 的频率录制为音频，并使用 Kaldi(Povey 等人，2014 年) 转录为文本，然后进行人工检查和纠正。访谈问题没有以音频格式记录，而是根据参与者的回答开发的。参与者被要求从大量问题中随机回答三个问题，并完成一份自评抑郁量表 (SDS) 问卷 (Zung，1965 年)。这份问卷由 20 个项目组成，总原始得分范围从 20 到 80。在临床实践中，这些得分通常会转换为 SDS 指数 (原始得分乘以 1.25) 进行标准化解释，结果范围为 25 到 100。为了进行二元分类，使用的临界值为 63 分；得分 63 分或更高表明患有抑郁症。在 162 名参与者中，132 名被归类为非抑郁症患者，30 名被归类为抑郁症患者。

安卓语料库是通过安卓项目 (陶等人，2023 年) 收集的，用于从参与者的语音中研究抑郁症。它包括 118 名意大利母语者，其中 116 人接受了临床访谈。有 116 个音频文件，我们将它们从 44.1 重新采样到 16kHz 。我们使用 Whisper(拉德福德等人，2023 年) 转录音频记录，捕捉每次访谈中采访者的问题和参与

者的回答。在这 116 个人中，64 人被诊断患有抑郁症，而 52 人是健康对照组。参与者是否患有抑郁症是由医学专业人员根据《精神疾病诊断与统计手册》第五版 (DSM-5)(协会等人，2013 年) 进行标记的。所报告症状的详细描述不可用。

每次访谈由不同数量的问答 (QA) 对组成，这些对中每个参与者的口头回答 $\left(A_j^a\right)$ 的持续时间也不同。我们的音频处理管道首先将每次访谈分割成单独的问答轮次。对于问答轮次中每个参与者的口头回答，其采样率为 16 kHz，音频被输入到 Wav2Vec2-XLSR-53 音频编码器中。该模型具有一个卷积层来处理音频，每 20 毫秒输出一系列局部声学特征向量。然后这些特征由模型的 Transformer 层进行处理，得到一系列帧级嵌入，每个嵌入有 1024 维。这个序列中的帧数直接对应于特定音频话语的持续时间。

为了为每个可变长度的音频话语创建固定大小的表示，我们对 A_j^a 的所有帧级嵌入应用时间平均池化。这个过程将整个 1024 维嵌入序列聚合为一个单一的固定维向量。然后将这个向量线性投影到 768 维，以确保在进行多模态融合之前与我们的文本特征维度对齐。

在模态融合和 D-CoPE 增强之后，我们为一次访谈中的所有问答 (QA) 轮次获得了一系列固定大小的表示，记为 $\{z'_1, z'_2, \ldots, z'_k\}$。为了在为对话级 Transformer 进行批处理时管理不同访谈中问答对数量 (k) 的变化，这些序列用零填充到统一的最大序列长度。这个最大长度是根据访谈长度的第 95 百分位数确定的，以问答对的数量来衡量。具体来说，对于 DAIC 和 DAIC 合成数据集，这个最大长度设置为 35，对于 EATD 数据集为 3，对于安卓语料库为 26。

D 评估指标

对于二分类任务，我们主要使用 F1 分数。真正例 (TP) 指被正确预测为患有抑郁症的参与者。假正例 (FP) 指被错误预测为患有抑郁症的参与者。假负例 (FN) 指实际上患有抑郁症但被模型错误分类为非抑郁症的参与者。

指标定义如下:

- 精确率衡量模型做出的所有正预测中真正例预测的比例。定义为:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

- 召回率 (也称为敏感度) 衡量所有实际正例中真正例预测的比例。定义为:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

- F1 分数是精确率和召回率的调和平均值，提供了两者的平衡度量。定义为:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

- AUC-ROC(受试者工作特征曲线下面积) 评估模型在所有可能的分类阈值下区分正类和负类的能力。AUC 越高，整体可分离性越强。

对于回归任务，我们报告基于误差的指标，这些指标捕获预测与真实值之间偏差的大小:

- 均方根误差 (RMSE) 是一种广泛使用的指标，用于衡量预测误差的平均大小。其计算方式为:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2} \tag{18}$$

其中 y_i 表示真实值，$\widehat{y}_i$ 是预测值，N 是测试集中样本的总数。RMSE 越低，预测准确性越高。

- 平均绝对误差 (MAE) 衡量预测值与实际值之间的平均绝对差异，提供与目标变量相同单位的可解释指标。其定义为:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i| \tag{19}$$

与 RMSE 相比，MAE 对大误差不太敏感，并提供了模型性能的补充视图。

E 其他实验结果

表 3 显示了每个数据集的实验结果，重点关注 AUC-ROC 和 MAE。

表 4 中呈现的跨数据集验证结果表明了几个关键观察结果。首先，在 DAIC-WOZ 数据集上训练的模型能够较好地迁移到 Androids 数据集，并在 EATD 数据集上表现尚可。然而，当模型在 EATD 上训练时，其在所有其他数据集上的性能始终较差。

模型	模态	DAIC-WOZ		EATD		安卓机器人		DAIC 合成数据	
		AUC-ROC	平均绝对误差	AUC-ROC	平均绝对误差	AUC-ROC	平均绝对误差	AUC-ROC	平均绝对误差
GCN-PE(布尔迪索等人，2024 年)	I	0.89	4.06	0.60	5.35	0.64	-	0.88	4.00
WU(吴等人，2023 年)	A + T	0.84	3.90	0.71	3.79	0.77	-	0.85	3.86
MMPF(杨等人，2024 年)	A + T	0.81	4.60	0.73	4.01	0.76	-	0.82	4.52
ACMA(约尔楚恩等人，2024 年)	A + T	0.76	4.64	0.69	4.61	0.72	-	0.74	4.32
语言模型 (张等人，2024 年)	A + T	0.80	4.54	0.70	4.21	0.74	-	0.83	4.15
CAMFM(薛等人，2024 年)	A + T	0.84	4.24	0.77	3.69	0.76	-	0.85	4.06
DAI(戴等人，2021 年)	I + A + T + V	0.84	4.20	0.72	3.90	0.77	-	0.83	4.18
HCAG(牛等人，2021 年)	I + A + T	0.85	4.31	0.70	4.98	0.75	-	0.79	3.99
SHEN(沈等人，2022 年)	I + A + T + V	0.80	4.72	0.77	4.13	0.72	-	0.80	4.63
MILI(米林采维奇等人，2023 年)	I + T	0.82	4.60	0.65	5.37	0.70	-	0.79	4.46
SEGA(陈等人，2024 年)	I + A + T + V	0.80	4.54	0.77	4.44	0.74	-	0.79	4.41
GCN(布尔迪索等人，2023 年)	I + T	0.84	4.46	0.69	4.97	0.72	-	0.82	4.38
AGAR(阿加瓦尔等人，2024 年)	I + T	0.80	5.36	0.66	4.98	0.75	-	0.80	5.32
对话变换器 (我们的)	I + A + T	0.89	3.47	0.75	3.64	0.80	-	0.90	3.46

表 3: 最佳结果用粗体显示，“-”表示无可用数据。AUC-ROC 越高，MAE 越低，性能越好。

训练与测试	DAIC-WOZ				EATD				安卓机器人				DAIC 合成数据			
	F1	均方根误差	曲线下面积-ROC	平均绝对误差	F1	均方根误差	曲线下面积-ROC	平均绝对误差	F1	均方根误差	曲线下面积-ROC	平均绝对误差	F1	均方根误差	曲线下面积-ROC	平均绝对误差
DAIC-WOZ	0.82	3.86	0.89	3.47	0.67	4.47	0.74	4.10	0.72	4.20	0.79	3.87	-	-	-	-
EATD	0.61	5.22	0.69	4.78	0.72	4.04	0.75	3.64	0.59	5.10	0.65	4.59	0.61	5.25	0.68	4.97
安卓机器人	0.70	-	0.77	-	0.66	-	0.72	-	0.73	-	0.80	-	0.69	-	0.76	-
DAIC 合成数据	-	-	-	-	0.68	4.40	0.74	3.96	0.74	4.10	0.78	3.66	0.83	3.84	0.90	3.46

表 4: 跨数据集验证结果。

此外，在安卓数据集上训练的模型能够有效地推广到 DAIC-WOZ 和 DAIC-Synthetic 数据集，不过在 EATD 数据集上测试时性能略有下降。相反，在 DAIC-Synthetic 数据集上训练的模型在 EATD 数据集上表现相当不错，在安卓数据集上取得了更好的结果。

这些发现表明，安卓数据集与 DAIC-WOZ 数据集具有更相似的半结构化访谈形式，而 EATD 数据集有限的问题多样性和较短的会话长度阻碍了模型从其他数据集学习可推广特征的能力。

由于合成数据主要用于扩充训练数据集，为避免过拟合和偏差风险，我们未在 DAIC-WOZ 和 DAIC-Synthetic 之间进行跨数据集验证。

F 超参数细节

在表 5 中，我们展示了框架中通用训练和所有其他模块的超参数。

G 基线实现

在本节中，我们提供了表 1 中列出的所有基线模型的实现细节。

- GCN-PE(Burdisso 等人，2024 年): 本研究构建了基于 LongBERT(Beltagy 等人，2020 年) 和图卷积网络 (GCN) 的模型，以分析参与者的回答和面试官的提示。他们证明，使用面试官提示的模型可以达到高精度，甚至达到了当前的最优性能。定性分析表明，这些模型倾向于关注访谈中特定的、局部的片段，特别是当面试官询问关于参与者心理健康史的针对性问题时。这表明模型学会将这些提示用作区分的捷径，而不是通过自身语言真正理解患者的抑郁状态。

- WU(Wu 等人，2023 年): 在本研究中，基础模型使用自监督学习 (SSL) 进行预训练，以解决基于语音的抑郁症检测 (SDD) 中的数据稀疏问题。主要方法包括分析来自基础模型 (如 wav2vec 2.0、HuBERT 和 WavLM) 各层的 SSL 表示，以识别抑郁症的指标。随后，这些模型在与自动语音识别 (ASR) 和自动情感识别 (AER) 相关的任务上进行微调，以促进知识向 SDD 的转移。

组件	超参数	探索的值
通用训练	优化器	亚当，亚当 W
	学习率	$\{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$
	批量大小	$\{4, \mathbf{8}, 16\}$
	轮次	$\{10, 20, 30\}$
	早停耐心值	$\{5, 10\}$
	权重衰减	$\{0, \mathbf{0.01}, 0.05, 0.1\}$
门控融合 (多层感知器)	层数	$\{1, 2\}$
	隐藏单元	$\{512, 768\}$
D - CoPE	层数	$\{1, 2\}$
	隐藏单元	{256, 512}
对话变换器	层数	$\{1, 2, 3\}$
	注意力头数	$\{4, \mathbf{8}, 12\}$
	前馈网络维度	$\{1536, 2048\}$
	随机失活率	$\{0.1, 0.15, 0.2, 0.25\}$
对抗分类器 (多层感知器)	层数	$\{1, 2, 3\}$
	隐藏单元	$\{64, 128, 256\}$
	拉姆达 λ	$\{0.01, 0.05, 0.1, \mathbf{0.25}, 0.5, 0.75, 1, 5\}$

表 5: 展示了为所提出的多模态抑郁症检测框架探索的超参数列表。以粗体突出显示的最优集在验证折上给出了最佳平均性能。

- MMPF(Yang 等人，2024 年): 提出了一个多模态融合框架，用于使用临床访谈中的音频、视频和文本流分析抑郁症。该系统从每个模态中提取特征，包括从段落向量中为选定回答提取的创新文本描述符以及从面部地标提取的视频描述符。这些特征通过深度卷积神经网络 (DCNN) 进行处理，以学习高级表示。然后将学习到的特征输入到深度神经网络 (DNN) 中，为每个模态预测初始的 PHQ - 8 抑郁症得分。最后，将各个管道的得分在融合 DNN 中进行组合，以产生最终的多模态 PHQ - 8 得分预测。

- ACMA(Iyortsuun 等人，2024 年): 这项工作将音频数据处理为梅尔频谱图和文本数据，文本数据由使用通用句子编码器编码的参与者回答表示。为此使用了两个单独的双向长短期记忆 (BiLSTM) 网络，每个网络后面都跟着一个注意力层，用于捕获重要的单模态特征。然后将这些处理后的单模态表示输入到 ACMA 网络中，该网络利用加法注意力机制对跨模态交互进行加权和组合，学习语音和文本线索之间的关系。

- LLM - 声学 (Zhang 等人，2024 年): 这项工作专注于三个主要步骤: 首先，从语音信号中提取离散声学地标。其次，使用低秩适应 (LoRA) 通过跨模态指令对模型进行微调。这一步骤教导语言模型理解这些地标及其与文本的关系，同时纳入关于说话者抑郁症状态的“提示”。最后，我们采用 P - 调优来训练语言模型，以整合文本和学习到的地标表示，用于抑郁症检测的最终任务。

- CAMFM(Xue 等人，2024 年): 这项工作提出了一个用于检测抑郁症的多模态模型，该模型将多个级别的音频特征与文本句子嵌入进行整合。对于音频组件，我们提取了低级描述符 (LLD)、梅尔频谱图和来自 wav2vec 模型的特征。然后使用多级音频特征交互模块 (MAFIM) 将这些特征进行组合，以形成全面的音频表示。在文本领域，我们利用预训练的 BERT 来获得句子级嵌入。然后使用基于通道注意力的新型多模态融合模块将这些不同的音频和文本表示融合在一起，从而实现异构数据的整合。

- DAI(Dai 等人，2021 年): 这项工作首先通过基于提取的主题进行上下文感知分析，从音频、视频和语义数据中构建一个高维特征向量。所提出方法的核心涉及一个两阶段特征选择算法。首先，一种过滤方法对高维特征进行排序，以识别候选信息子集。接下来，一种包装方法通过顺序添加特征并利用支持向量机 (SVM) 模型来细化该子集，以仅保留那些提高预测准确性的特征。

- HCAG(Niu 等人，2021 年): 本文提出了一种用于抑郁症检测的分层上下文感知图注意力模型。该模型首先利用带有门控循环单元 (GRU) 的顺序编码器和加法注意力机制，从文本 (使用 GloVe 嵌入) 或音频 (利用 MFCC 和 eGeMAPS 特征) 中为每个问答对生成表示。在此之后，一个主题级上下文编码器构建一个图，其中问答对作为节点。然后使用图注意力网络 (GAT) 聚合上下文信息，并在定义的上下文窗口内学习这些问答对之间的关系。

- SHEN(Shen 等人, 2022 年): 从音频记录中提取梅尔频谱图, 并使用 NetVLAD(Arandjelovic 等人, 2016 年) 将其转换为固定长度的音频嵌入。然后将这些嵌入输入到门控循环单元 (GRU) 网络中。同时，使用 ELMo 从访谈转录本生成的句子嵌入由配备注意力机制以捕获重要语言线索的双向长短期记忆 (BiLSTM) 网络进行处理。文本 (BiLSTM) 和音频 (GRU) 分支的特征表示进行连接，并应用“模态注意力”机制来权衡它们各自的贡献。然后将得到的融合表示传递到全连接网络进行抑郁症的最终二元分类。

- MILI(Milintsevich 等人，2023 年): 这项工作提出了一个使用分层架构处理文本转录本以检测抑郁症的模型。首先，一个句子 - RoBERTa 模型对各个对话轮次进行编码。然后，一个带有加法注意力机制的双向长短期记忆 (BiL - STM) 处理这些轮次嵌入，以创建用于预测的访谈的综合表示。

- 世嘉 (陈等人，2024 年): 这项工作专注于通过构建结构元素图，在抑郁症评估中运用专家知识。它建立了一个有向无环图，在每个访谈回合中，信息从辅助节点 (音频、视频和问题) 流向中心节点 (答案记录)。中心记录节点相互连接以捕捉时间依赖性，并且所有中心节点都连接到一个表示整个访谈语义的汇总节点。最终，构建了一个图注意力网络，以便从构建的图中学习。

- 图卷积网络 (布尔迪索等人，2023 年): 这项工作使用图卷积网络 (GCN) 对访谈记录进行分类。改进后的 GCN 为边，特别是自连接，采用了一种新的加权方案，其中权重由 PageRank 算法确定，以反映每个节点 (单词或文档) 的重要性。从单词节点 (独热向量) 和文档节点 (TF-IDF 特征) 创建一个异构图。连接基于单词-单词链接的点互信息和单词-文档链接的 TF-IDF。这种方法能够对长距离语义进行建模，并有效地将受试者分类为抑郁或对照组。

问题功能	定义	示例	来源
“开放式”	鼓励参与者就某个主题自由广泛地表达自己，包括他们的经历、想法或感受的问题，通常不仅仅是寻求简短或特定答案。	“你喜欢做什么娱乐活动?”	动机访谈 (米勒和罗尔尼克，2012 年)，认知访谈 (萨默斯 - 弗拉纳根和萨默斯 - 弗拉纳根，2012 年)
“改变谈话”	旨在帮助参与者表达与行为、认知或情境改变相关的动机、原因、愿望、能力或需求的问题。	“是什么促使你寻求帮助?”	动机访谈 (米勒和罗尔尼克, 2012 年)，阶段匹配理论 (普罗查斯卡和韦利泽，1997 年)
“中性信息收集”	寻求特定事实细节、澄清客观信息或对具体询问直接回答的问题，通常期望得到有限或简短的回答。	“你获得良好夜间睡眠的难度有多大?”	话语分析标记集 (科尔和艾伦，1997 年)，认知访谈 (萨默斯 - 弗拉纳根和萨默斯 - 弗拉纳根，2012 年)
“过渡性”	访谈者用于组织对话、管理话题之间的过渡、引入或结束部分、总结要点或管理访谈流程的话语 (问题或陈述)。	“你经常有那种感觉吗?”	动机访谈 (米勒和罗尔尼克，2012 年)
“具体追问”	后续问题，旨在引出关于参与者先前介绍的主题或陈述的更详细信息 (阐述或具体示例)，通常以中性、非引导性的方式进行。	“你被诊断出患有抑郁症吗?”	认知访谈 (萨默斯 - 弗拉纳根和萨默斯 - 弗拉纳根，2012 年)
“支持性”	主要传达同理心、理解、对参与者感受或经历的认可、提供鼓励、建立融洽关系或肯定其优势的陈述或问题。	“听起来极其困难。”	动机访谈 (米勒和罗尔尼克，2012 年)，共情关系疗法 (韦利维塔和普，2020 年)
“其他”	不属于任何其他已定义功能类别的话语。这可能包括非常简短的反馈、不完整或中断的句子、难以理解的言语或与结构无关的离题言论。	“是的。”; “嗯。”	话语分析标记集 (科尔和艾伦，1997 年)

表 6: 我们根据既定理论定义的所有面试问题功能 (IQF) 的完整列表。MI 代表动机性访谈，CI 代表临床访谈技巧，TMC 代表跨理论变化模型，DAMSL 指多层对话行为标记，ERT 代表共情反应分类法。

- AGAR(Agarwal 等人，2024 年): 这项工作提出了一种多视图架构，旨在通过明确考虑话语结构来改进从患者-治疗师访谈记录中进行的自动抑郁评估。核心方法包括将记录分为两个不同的“视图”——一个用于治疗师问题，另一个用于患者回答。每个视图由一个专用的视图编码器处理，该编码器利用多头注意力从句子变换器生成的句子级编码中学习特定表示。重要的是，这些视图编码器通过交叉注意力机制进行交互，使它们能够通过共享注意力分数以相互依赖的方式学习。

H 提示标注

在表 6 中，我们列出了根据传播研究和临床心理治疗中的既定方法和理论定义的所有问题功能 (QF)。

由于资源有限，我们使用了三个大语言模型进行标注:GPT-3.5 Turbo、GPT-4o 和 LLaMA3-70B。连同表 7 中总结的提示模板，每个大语言模型都被提供了每个 QF 的定义和示例以及相关上下文。当前需要标注的句子的“上下文”由前面的对话组成，上下文长度经过精心定义以适应我们的资源限制；在我们的研究中，上下文长度设置为 5。因此，对于每个问答对中每个面试官的话语，大语言模型最多考虑前面 5 个话语的信息来标注当前话语。

提示模板
两位参与者，即面试官和受访者，正在进行临床访谈。对话围绕 {背景} 展开。现在面试官 {面试官} 说 :{当前句子}。根据选项 [开放式、转换话题、中性信息收集、过渡性、具体探究、支持性、其他] 预测句子 { 当前句子 } 的问题功能，考虑对话背景，无需解释，仅输出 [开放式、转换话题、中性信息收集、过渡性、具体探究、支持性、其他] 中的标签。

表 7: 用于注释访谈者表达的每个话语的问题功能的提示。

为了确定最终标签，我们对三个模型的输出进行了多数投票。如果模型之间没有达成共识，我们将该话语标记为“其他”。

I 基础模型

我们的多模态抑郁症检测框架的成功在很大程度上依赖于从文本和音频模态中获得的初始特征表示的质量。本节概述了我们选择编码器背后的推理。

I.1 文本特征提取

为了处理面试官问题 $\left(Q_j^{\text{text}}\right)$ 和参与者转录的回答 $\left(A_j^{\text{text}}\right)$,我们使用 XLM-RoBERTa(XLMR)(Conneau 等人，2019 年；Liu 等人，2019 年) 作为我们的文本编码器。关键原因是其强大的多语言能力。我们的数据集涵盖多种语言，包括英语 (DAIC-WOZ、DAIC-Synthetic)、汉语普通话 (EATD) 和意大利语 (Androids)。XLMR 使用大规模多语言语料库在 100 种语言上进行预训练，并基于强大的 RoBERTa 架构构建，使其能够生成高质量、跨语言一致的嵌入。这种选择使我们能够在不使用单独的特定语言模型的情况下，保持跨数据集的方法一致性。

我们考虑了几种选择:

- 单语模型 (例如，用于英语的 BERT-base(Devlin 等人，2018 年)) 每种语言都需要不同的编码器，这会使系统设计复杂化，并可能引入跨语言的不一致性。
- 诸如 mBERT(Devlin 等人，2018 年) 等较老的多语言模型通常在性能上不如 XLM-R，因为其预训练策略和语料库规模更优越。
- 大型生成式语言模型 (例如，GPT 系列) 展现出卓越的语言理解能力 (赫斯特等人，2024 年)，但将它们用作特征编码器——尤其是进行微调时——对于在我们的顺序管道中处理大量短文本片段而言，计算成本高昂且不切实际。

因此，XLMR 在多语言表示能力和计算效率之间提供了良好的权衡，使其非常适合在我们的下游模块中进行句子级嵌入提取。

I.2 音频特征提取

为了对参与者的口语回答 $\left(A_j^{\mathrm{a}}\right)$ 进行编码，我们使用了 Wav2Vec2-XLSR-53(XLSR-53)(Conneau 等人，2020 年；Baevski 等人，2020 年)。这个选择基于几个因素:

- 多语言能力。XL-SR53 在包括英语、汉语普通话和意大利语在内的 53 种语言的语音上进行了预训练。它对原始音频波形采用自监督学习方法，因此在无需特定语言微调的情况下，能特别有效地在不同语言数据集中生成一致且可比的音频表示。
- 从原始音频中学习。与依赖手工制作特征 (如 MFCCs 或 eGeMAPS(Ey-ben 等人，2010 年，2015 年)) 的传统声学方法不同，基于 Wav2Vec2 的模型直接从原始波形中学习表示。这种方法与我们尽可能开发数据驱动、端到端可训练系统的目标一致。

-性能。Wav2Vec2 及其多语言变体 XLSR-53 在广泛的语音处理基准测试中都取得了出色的成绩, 凸显了所学习音频表示的质量和通用性。其他自监督语音模型，如 HuBER(Hsu 等人，2021 年) 和 WavLM(Chen 等人，2022 年)，也是有力的候选者。然而，XLSR-53 对跨语言能力的明确关注以及它在多语言语音任务中的实证成功使其特别适合我们的多语言设置。