基于大语言模型提示的睡眠健康与生活方式数据集中睡眠障碍自动分类研究

**摘要**

睡眠障碍是一个紧迫的全球公共问题，影响着全球约 30% 的人口。本研究利用LLMs 强大的语义理解与知识推理能力，对涵盖睡眠、生活方式及相关健康因素的数据集进行睡眠障碍分类。通过三种新颖的提示范式来引导LLMs 自动完成分类器的设计，训练与评估。实验结果表明，使用分解提示自动找到的支持向量机分类器，分类准确率高达 91.9%（F1 - 分数：0.919），在准确率上显著优于传统的零样本提示和多样本提示方法。这项研究独特地将大语言模型的语义理解知识推理能力与自动化机器学习无缝地融合，为健康信息学中的睡眠障碍分类提供了一种新范式。

**关键词**：睡眠障碍分类；大语言模型；提示工程；健康信息学

### 引言

### 1.1 睡眠障碍问题的严峻性

睡眠障碍是一个重大问题，影响着很大一部分人群，并可能对整体健康和生活质量产生重大影响。根据世界卫生组织（WHO）的统计，全球约有10%的人口受到睡眠障碍的困扰，且这一比例在城市化进程加快的背景下呈上升趋势。睡眠障碍不仅会导致日间疲劳、注意力下降，还可能引发心血管疾病、糖尿病等慢性疾病。因此，准确识别和分类睡眠障碍对于改善公众健康具有重要意义。

### 1.2 传统睡眠障碍分类方法的困境

在过去，睡眠障碍分类主要依赖于传统机器学习算法，如决策树算法[1]、支持向量机（SVM）[2]、随机森林算法[3]等。这些算法在睡眠障碍研究中发挥了一定作用。研究人员利用决策树算法对睡眠监测数据进行分析，通过一系列的条件判断和分支决策，试图识别出不同睡眠障碍的特征模式 。支持向量机则通过寻找一个最优的分类超平面，将正常睡眠数据和睡眠障碍数据区分开来。随机森林算法通过构建多个决策树并综合它们的预测结果，提高了分类的准确性和稳定性 。

然而，传统机器学习算法在睡眠障碍分类中存在诸多困境。这些算法的每一个步骤都需要大量的手动操作。在数据预处理阶段，研究人员需要手动处理数据缺失值、异常值，对数据进行标准化、归一化等操作，以确保数据的质量和可用性。在模型训练和调优过程中，研究人员需要手动选择合适的算法、设置模型参数，并通过反复试验来优化模型性能 。

这种手动操作的方式不仅耗费大量的人力和时间成本，还容易受到人为因素的影响，导致结果的准确性和可靠性存在一定的局限性。而且，传统机器学习算法的应用高度依赖专家经验。专家需要根据自己的专业知识和经验，选择合适的算法、确定特征工程的方法以及调整模型参数。对于复杂的睡眠障碍分类问题，不同专家的经验和判断可能存在差异，导致分类结果的不一致性 。

传统机器学习算法在睡眠障碍分类中存在的这些问题，限制了睡眠障碍研究的发展和应用，需一种新的技术和方法来突破这些困境。

### 1.3 大语言模型带来的新契机

大语言模型（LLMs）作为自然语言处理领域的前沿技术，近年来取得了显著的进展。它基于 Transformer[14] 架构，通过在海量文本数据上进行无监督预训练，学习到了丰富的语言知识。它在众多领域展现出了巨大的潜力，为解决复杂问题提供了新的思路和方法 。

在医学研究领域[4]、[6]、[7]、[8]、[9]，大语言模型的应用为睡眠障碍分类带来了新的契机。大语言模型能够理解和处理自然语言，这使得它可以直接对睡眠健康与生活方式数据集中的文本信息进行分析和解读。它可以从患者的睡眠日志、生活习惯描述等文本数据中提取关键信息。

大语言模型还具备强大的知识推理能力[5]，它可以结合已有的医学知识和睡眠障碍的诊断标准，对提取到的信息进行综合分析和判断，从而实现对睡眠障碍的自动分类。在面对复杂的睡眠障碍症状时，大语言模型能够通过推理和判断，准确地识别出不同类型的睡眠障碍，如失眠、睡眠呼吸暂停等。

而且，大语言模型还可以通过对大量数据的学习，发现睡眠障碍与其他因素之间的潜在关系，为睡眠障碍的诊断和治疗提供新的依据 。

### 1.4 研究贡献

在睡眠障碍分类研究领域，传统方法高度依赖手动操作和专家经验，这不仅效率低下，还难以应对日益复杂的睡眠健康数据，存在明显局限性。

而提示策略（Prompting）作为大语言模型应用中的关键技术，对引导模型生成预期输出起着决定性作用。不同的提示策略在各类任务中的性能表现参差不齐，特别是如何巧妙运用提示策略，使大语言模型仅依据文本信息就能够准确从睡眠健康与生活方式数据中提取关键信息，实现自动睡眠障碍分类，这一难题亟待攻克。本研究在该背景下，具有多方面重要且独特的贡献：

* **创新模型应用**：开创性地将大语言模型引入自动睡眠障碍分类领域。传统的睡眠障碍分类方法面对海量、复杂且多变的睡眠健康与生活方式数据时，精度不高、适应性差等问题愈发凸显。而大语言模型凭借其强大的语言理解和生成能力，突破了传统方法的重重局限。本研究构建了一套全新的、基于大语言模型的自动睡眠障碍分类方法，这一方法的核心优势在于，只要提供文本信息，就能让模型自动进行机器学习（auto ML），实现对睡眠障碍的高精度预测，为睡眠障碍的准确诊断开辟了一条崭新且高效的途径。

1. **深度挖掘提示策略助力自动分类：**本研究对零样本提示（Zero-shot Prompting）和少样本提示（Few-shot Prompting），分解提示（Decomposed Prompting）技术在大语言模型中的应用展开了深度探索。本研究系统地挖掘了这些提示策略在自动睡眠障碍分类任务中的全新潜力，清晰明确了它们在基于文本的睡眠健康分类中的优势与不足。这些深入分析为后续提示策略的优化提供了明确的方向，有助于持续提升大语言模型在仅依据文本信息进行自动睡眠障碍分类时的性能。

* **引领跨学科融合与自动化创新：** 本研究通过提示驱动大语言模型，实现了仅凭借文本信息就能自动进行机器学习分类的创新应用，这打破了传统机器学习算法的束缚，真正实现了睡眠障碍分类的自动化和智能化。这种创新应用不仅有效解决了睡眠医学领域的实际问题，更重要的是，它为大语言模型在医疗健康领域的更广泛应用积累了宝贵经验，极大地推动了跨学科技术的深度融合与发展，为睡眠医学和人工智能的交叉研究开辟了一条充满无限可能的新路径。
* **拓展自动化应用场景：** 在实际应用方面，大语言模型展现出了强大的自动化拓展能力。它可与多种技术相结合，进一步拓展其在睡眠障碍分类及相关领域（传感器技术、可穿戴设备等）的自动化应用价值。

本研究的成果不仅有助于显著提高睡眠障碍的诊断和治疗水平，改善患者的生活质量，而且随着技术的不断发展和完善，大语言模型在睡眠医学领域基于文本信息的自动分类应用前景也将更加广阔，为推动整个睡眠医学行业的自动化、智能化发展注入了强大动力。

## 三、睡眠健康与生活方式数据集

### 3.1 数据集来源、构成

本研究使用的睡眠健康与生活方式数据集来源于 Kaggle 网站[10]。本次使用的睡眠健康与生活方式数据集包含 374 行 13 列数据。具体如下：

1. **个人基本信息**：Person ID（个人编号）作为每个受访者的唯一标识符，有助于在数据处理和分析过程中准确识别和跟踪个体数据。、Gender（性别）信息可以用于研究不同性别在睡眠障碍发生率和睡眠模式上的差异。Age（年龄）是影响睡眠的重要因素之一，随着年龄的增长，睡眠质量往往会下降，睡眠障碍的发生率也会增加 。Occupation（职业）则反映了工作性质、工作时间和工作压力等因素对睡眠的潜在影响。
2. **睡眠相关特征**：Sleep Duration（睡眠时长）直接反映了个体的睡眠时间，充足的睡眠时长对于维持身体健康和正常的生理功能至关重要。Quality of Sleep（睡眠质量评分）则是一个主观评价指标，通过量表（1 - 10）来衡量，它反映了睡眠的深度、连续性和恢复效果等方面 。
3. **生活方式因素**：Physical Activity Level（身体活动水平）反映了个体的日常运动量。Stress Level（压力水平）是影响睡眠的重要因素之一 。BMI Category（BMI 类别）是衡量个体体重状况是否健康的指标，与睡眠障碍密切相关。Daily Steps（每日步数）则是一种简单直观的衡量身体活动水平的指标，可以了解个体的日常活动量，进而分析其对睡眠的影响 。
4. **健康指标**：Blood Pressure（血压）对于维持身体各器官的正常功能至关重要。Heart Rate（心率）反映了心脏的功能状态。

这些变量相互关联，共同反映了受访者的睡眠健康状况和生活方式特点。通过对这些变量的深入分析，可以揭示Sleep Disorder（睡眠障碍情况）与生活方式之间的潜在关系，为睡眠障碍的诊断、治疗和预防提供科学依据 。

### 3.2 数据特征分析与可视化

**一、数值型特征的统计信息（图 1）**

从这些数值型特征的统计数据中可以推测：

* **睡眠时长**：平均睡眠时长约为 7.13 小时，标准差较小，说明整体分布相对集中，大部分人的睡眠时长在 6.4 - 7.8 小时之间，符合成年人正常睡眠时长范围。但仍有部分个体可能存在睡眠时长不足或过长的情况，可能与生活习惯、工作压力等因素有关。
* **睡眠质量评分**：平均评分为 7.31 分，标准差 1.20 分，个体之间存在一定差异。25% - 75% 分位数显示大部分人的评分在 6 - 8 分之间，整体睡眠质量处于中等偏上水平。不过，仍有相当一部分受访者睡眠质量较差，评分低于 5 分，这可能受到多种因素的综合影响，如心理压力、生活环境等。
* **身体活动水平**：平均水平为 59.17，标准差 20.83 相对较大，说明不同个体之间的身体活动水平差异明显。这与个人运动习惯、职业特点等因素密切相关，例如从事体力劳动的职业人群和久坐办公室的人群身体活动水平可能有很大差异。
* **压力水平**：平均压力水平为 5.39，标准差 1.77，压力水平的分布有一定离散性。不同个体面临的压力程度不同，可能与工作性质、生活事件等因素有关。较高的压力水平可能会对睡眠质量等健康指标产生负面影响。
* **心率**：平均心率 70.17 次 / 分钟，标准差 4.14，心率的波动范围相对较小，大部分人心率在 68 - 72 次 / 分钟之间。这反映出该数据集中心率总体较为稳定，但仍有部分个体的心率可能偏离正常范围，可能与身体健康状况、运动情况等有关。
* **年龄**：平均年龄为 40.06 岁，标准差 13.34，说明年龄分布有一定的离散性。不同年龄段的人群在睡眠相关特征上可能存在差异，例如年龄较大者可能睡眠时长较短或睡眠质量较差。
* **每日步数**：平均每日步数为 7070.26 步，标准差 3344.52，步数的差异反映了不同个体的运动习惯和活动量不同。步数可能与身体活动水平、睡眠质量等存在关联，步数较多者可能身体活动水平较高，睡眠质量也较好。

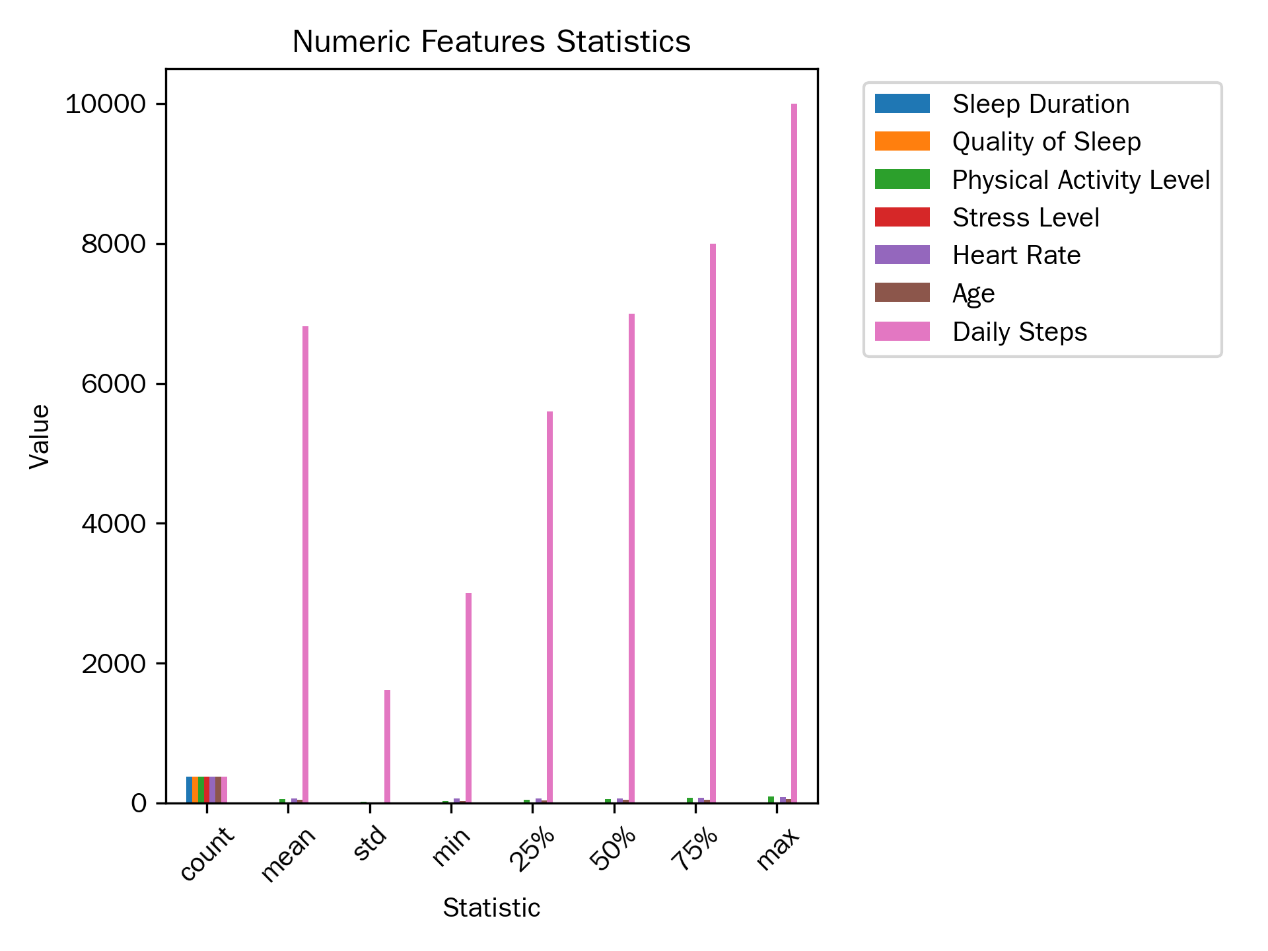


图 1：数值型特征统计图表

1. **分类型特征的分布信息（图 2）**

从分类型特征的分布中我们可以看出：

* **性别**：男性有 212 人，女性有 162 人，男性数量略多于女性。不同性别的人群在睡眠相关特征上可能存在差异，例如女性可能因生理因素在睡眠质量、压力水平等方面表现不同。
* **职业**：各个职业的人数相同，均为 46 人，反映出样本在职业方面具有一定的均衡性。不同职业可能由于工作压力、工作时间等因素，在睡眠时长、质量、压力水平等方面表现出不同的特征。例如，护士可能由于工作的轮班性质，睡眠时长和质量受到影响，压力水平较高。
* **BMI 类别**：正常体重的受访者有 156 人，超重的有 121 人，肥胖的有 97 人，正常体重和超重的受访者占比较高。不同 BMI 类别的人群可能在睡眠障碍的发生率上存在差异，肥胖人群可能更容易出现睡眠呼吸暂停等问题。
* **睡眠障碍情况**：睡眠正常的受访者有 236 人，有睡眠呼吸暂停的有 85 人，失眠的有 53 人。了解不同睡眠障碍类型的分布，有助于针对性地研究睡眠障碍的成因和预防措施。如睡眠呼吸暂停患者可能与肥胖、年龄等因素相关。

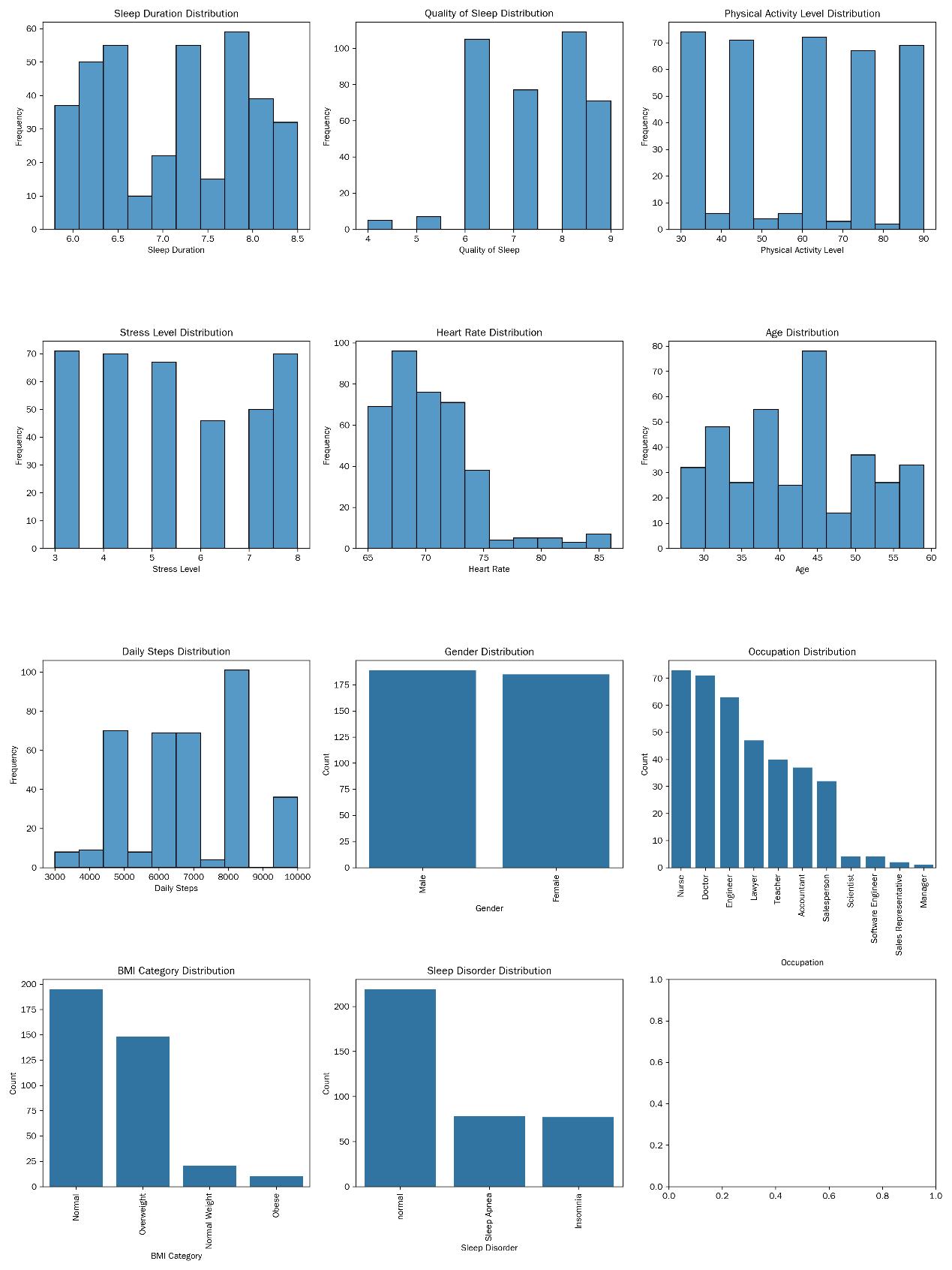


图 2：分类型特征分布图表

**三、特征间相关性分析**

计算了各特征之间的皮尔逊相关系数（对于分类型特征，采用合适的方法分析其与其他特征的关系，如卡方检验分析Gender与Sleep Disorder的关系等），通过绘制热力图展示相关系数矩阵（图 3）。

从相关系数矩阵中可以得出以下结论：

1. **睡眠时长与睡眠质量**：二者的相关系数为 0.883213，呈现出较强的正相关关系。这符合一般认知，即睡眠时长越长，身体和大脑能得到更充分的恢复，从而提高睡眠质量。这一结果提示在改善睡眠质量的措施中，可以考虑通过调整睡眠时长来实现。
2. **身体活动水平与其他特征**：身体活动水平与睡眠时长、睡眠质量的相关系数分别为 0.212360 和 0.192896，呈现较弱的正相关关系。这表明适度的身体活动对睡眠有一定的积极影响，但这种影响相对有限。身体活动水平与压力水平的相关系数几乎为 0，说明两者之间线性关系不明显。而与心率的相关系数为 0.136971，有较弱的正相关，可能是身体活动水平较高时会引起心率一定程度的上升。
3. **压力水平与其他特征**：压力水平与睡眠时长、睡眠质量分别呈现 - 0.811023 和 - 0.898752 的强负相关关系。这表明压力是影响睡眠的重要因素，当人们处于高压力状态时，身体分泌的应激激素会干扰入睡和睡眠质量。压力水平与心率的相关系数为 0.670026，呈正相关关系，说明压力越大，心率可能越高，反映出压力对身体生理指标的影响。
4. **心率与其他特征**：心率与睡眠时长、睡眠质量呈负相关关系，意味着心率较高时，睡眠时长和质量可能较差。这可能是由于心率异常反映了身体的某种不适状态，进而影响睡眠。
5. **年龄与其他特征**：年龄与睡眠时长、睡眠质量呈负相关关系，说明随着年龄的增长，睡眠时长可能减少，睡眠质量可能变差。年龄与压力水平呈正相关，可能年龄较大者面临的生活压力相对更大。
6. **每日步数与其他特征**：每日步数与睡眠时长、睡眠质量呈正相关关系，表明步数较多可能有助于延长睡眠时长和提高睡眠质量。每日步数与身体活动水平呈较强的正相关，符合预期，步数越多身体活动水平越高。

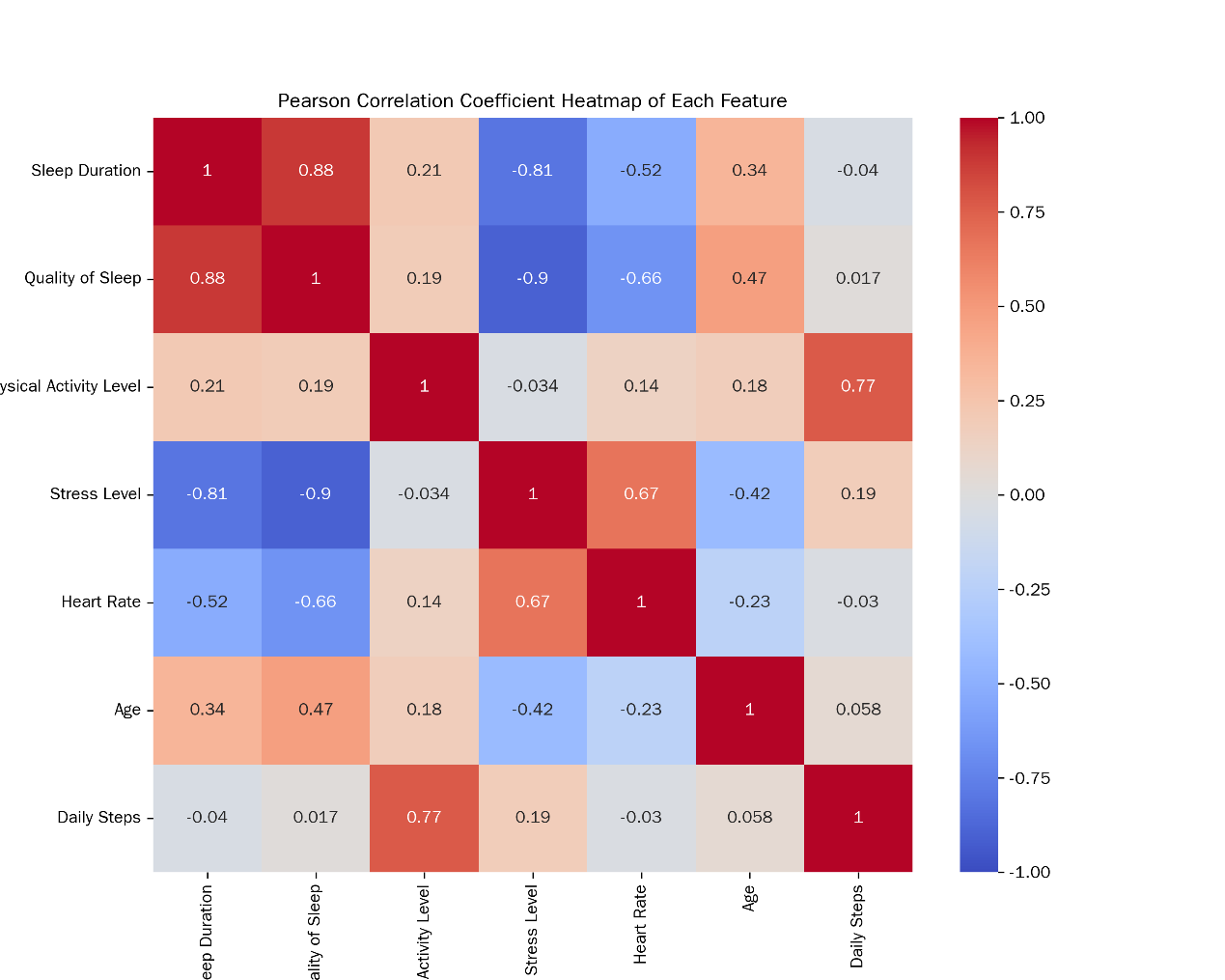


图 3：各特征之间皮尔逊相关系数热力图

通过对数据集中各特征的分布情况进行统计分析和可视化展示，我们可以更深入地了解数据的特点和规律，为后续利用大语言模型进行睡眠障碍分类提供有力的数据支持。在后续的研究中，我们可以根据这些分析结果，选择合适的特征和模型，提高睡眠障碍分类的准确性和可靠性 。

## 四、研究方法

### 4.1 大语言模型选择

在睡眠障碍自动分类这一研究领域，大语言模型的选型是影响研究成果准确性与效率的关键因素。经过综合考量，本研究最终选定豆包模型，主要基于以下几方面的深度剖析。

* 豆包模型构建于 Transformer[14] 架构之上，通过在海量的文本数据中进行无监督预训练，广泛且深入地学习了语言在语义、语法、语用等多方面的知识体系。其独特的多头注意力机制，能够并行地关注输入文本的不同部分，从而捕捉到文本中丰富的语义关联和上下文信息；多层神经网络结构则进一步对这些信息进行深度加工与特征提取，使得豆包模型能够从词汇、语句、篇章等多个维度深入剖析文本内容。
* 在自然语言处理任务中，豆包模型能充分理解睡眠健康领域专业的医学术语，如 “睡眠呼吸暂停低通气指数”“周期性肢体运动障碍” 等，并且，凭借良好的泛化性能，豆包模型可以快速适应睡眠健康领域复杂多变的文本数据，包括医疗记录、睡眠监测设备的日志数据等，这些数据来源广泛，格式和语言风格差异较大，但豆包模型都能有效处理。
* 此外，豆包模型还具备一些独特的优势，极大地助力了睡眠障碍自动化分类的实现。在交互层面，部分版本（如 PC 版）的豆包模型拥有友好的用户接口，能够直接接受分析 CSV 文件等常见的数据格式，这使得研究人员可以便捷地将睡眠健康与生活方式数据集导入模型进行处理，无需花费大量时间进行数据格式转换等预处理工作。在技术实现层面，豆包模型具备自动生成代码和执行代码的能力，在睡眠障碍自动化分类流程中，该能力可依据文本分析结果快速生成实现分类算法的代码逻辑，并直接执行，不仅大大提高了分类效率，还能够修正人工编写代码过程中可能出现的语法错误、逻辑漏洞等问题，为实现高效、准确的睡眠障碍自动化分类提供了有力支持。

综上所述，豆包模型凭借其强大的自然语言处理能力、对睡眠健康领域数据的高度适配性以及独特的交互和技术实现优势，成为本研究中睡眠障碍自动分类的不二之选。

### 4.2 精妙提示设计策略

如表1，提示细节， 我们设计了三种提示策略：

* + 零样本提示（Zero-shot Prompting）规则引导的分类探索: 核心概念是在不提供任何具体示例的情况下，仅凭借对任务的清晰描述和模型自身的预训练知识，引导模型生成相应的回答或完成特定任务 。这一策略的原理基于大语言模型在大规模数据上的预训练，使其具备了广泛的语言理解和知识储备能力，能够根据任务描述中的语义信息，从已学习的知识中提取相关内容并进行推理和判断,
* 少样本提示（Few-shot Prompting）数据模式驱动的分类优化: 其设计思路基于对数据模式的深入挖掘和利用，旨在通过提供一定数量的样本数据，引导大语言模型学习数据中的特征与分类结果之间的对应关系，从而实现更准确的分类. 90 样本提示的优势在于它能够利用训练集中的数据模式，为大语言模型提供更丰富的信息，从而提高分类的准确性 。
* 分解提示（Decomposed Prompting 任务拆解的分类进阶: 其核心在于将复杂的任务进行细致的拆解，转化为一系列更易于处理的子任务，从而引导大语言模型更高效地完成任务。在自然语言处理领域，这种策略能够充分发挥大语言模型的优势，提升任务完成的质量和效率.

每种策略包含两个子任务，任务一是对指定CSV文件中的数据进行多元分类，并生成新的CSV文件. 任务二（三种提示相同）是使用特定的评估指标对分类结果进行评估，并绘制相关图表。

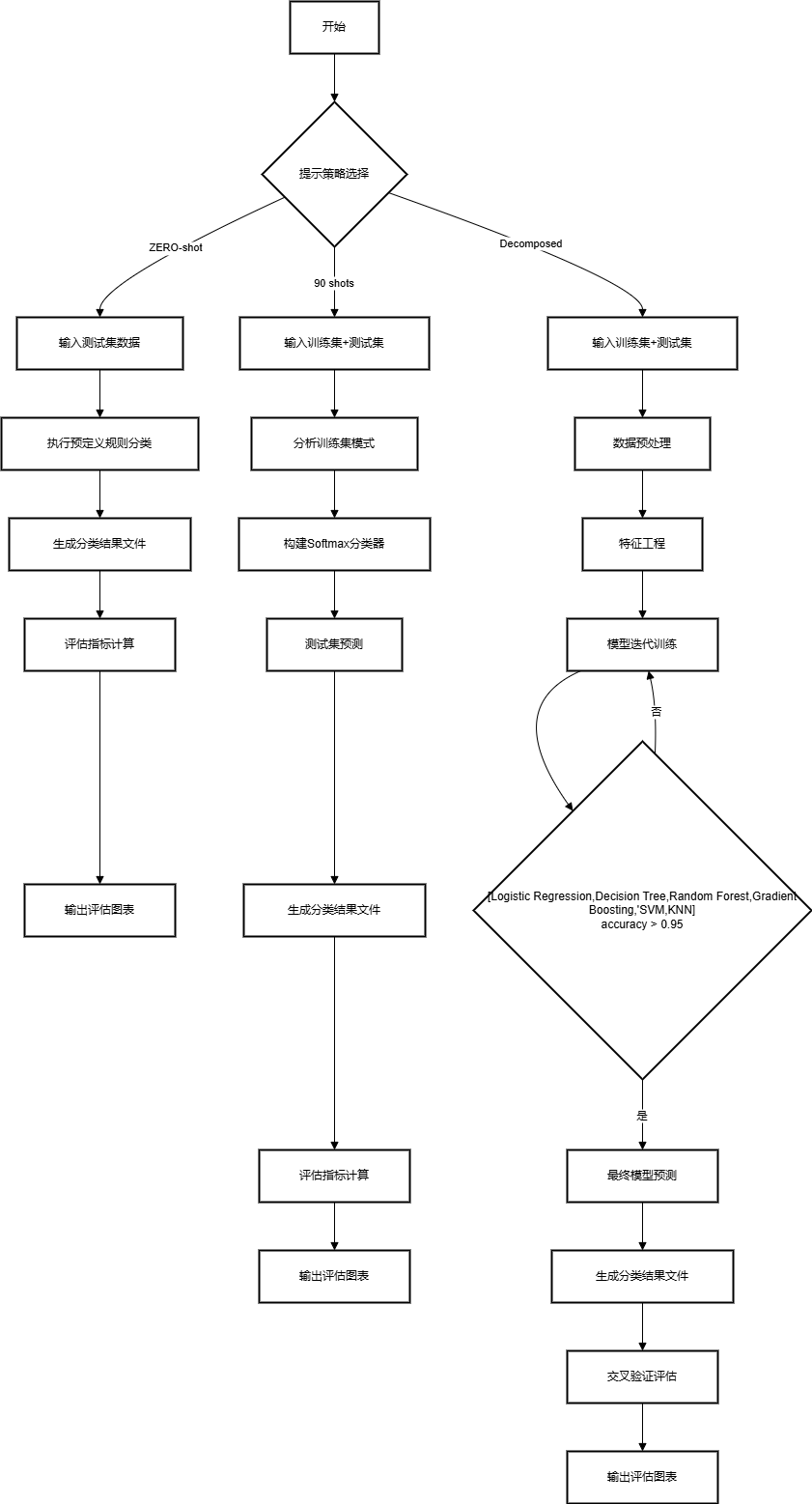


图4：提示策略流程图

## 五、实验设计与实施

### 5.1 实验环境搭建

实验环境的搭建是确保研究顺利进行的基础，其硬件和软件配置对实验结果的准确性和可靠性有着重要影响。在本次睡眠障碍分类研究中，实验选用的处理器为 Intel (R) Core (TM) i5 - 6400T CPU @ 2.20GHz，该处理器具备 2201 Mhz 的主频，拥有 4 个内核和 4 个逻辑处理器 。操作系统采用的是 Microsoft Windows 10 家庭中文版，版本为 10.0.19045 内部版本 19045 。大语言模型选用的是豆包PC版本 1.41.6 。

### 5.2 实验步骤流程

1. 数据集的划分及样本的选取：
   * 从原数据集中按三种类别（正常、睡眠呼吸暂停、失眠）随机各选取 30 个样本，共 90 个样本作为prompts 90 examples 。这一随机选取的方式确保了样本的代表性和随机性，能够在一定程度上反映数据集中不同睡眠障碍类型的特征 。将这 90 个样本保存为Sleep\_health\_and\_lifestyle\_dataset\_selected\_90.csv作为训练集
   * 在原数据集文件中删去以上 90 个样本，保存为Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv，作为真实标签（GROUND Truth）。真实标签是评估模型分类准确性的重要依据，通过将模型的预测结果与真实标签进行对比，可以准确地评估模型的性能
   * 将Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv复制一份，并删去最后一列（Sleep Disorder），保存为Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv作为测试集 。测试集用于评估模型在未见过的数据上的表现，能够检验模型的泛化能力和分类准确性
2. 手动上传以上 3 个csv文件到豆包 。这一步骤确保了大语言模型能够获取到所需的数据，为后续的分类器设计、训练和评估提供数据基础 。
3. 根据图4（详情见附件一）不同提示策略进行分类器设计、训练和评估

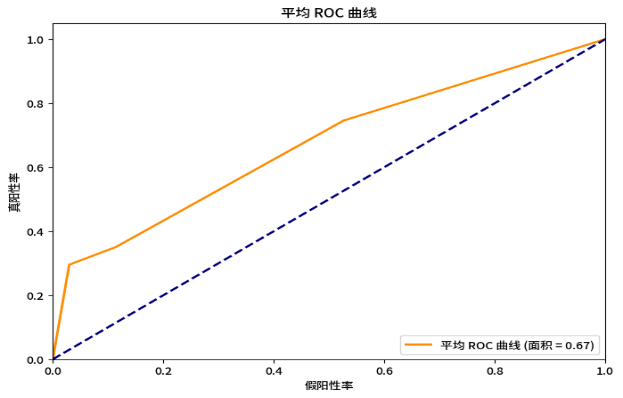
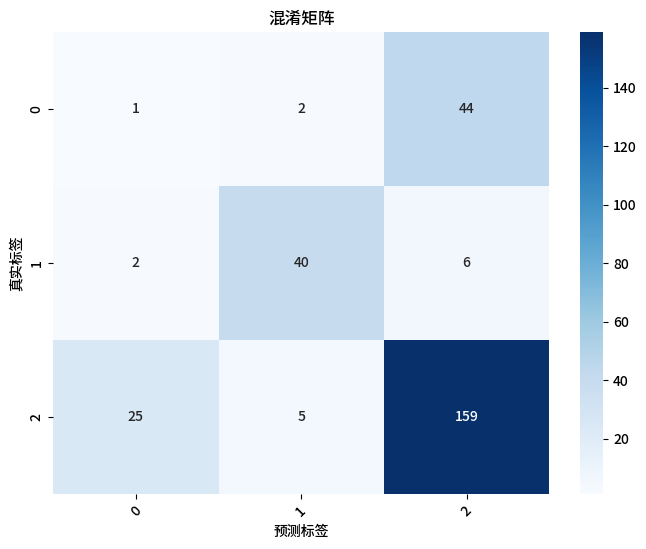
。

### 5.3 实验结果

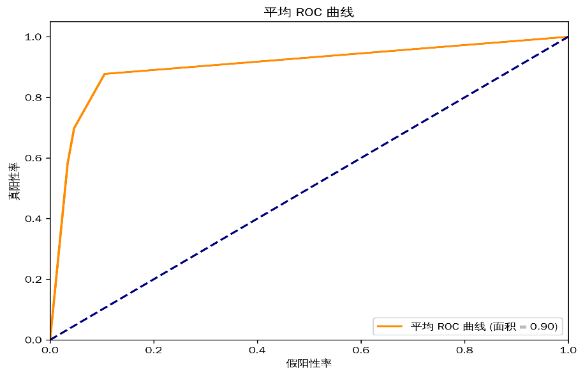
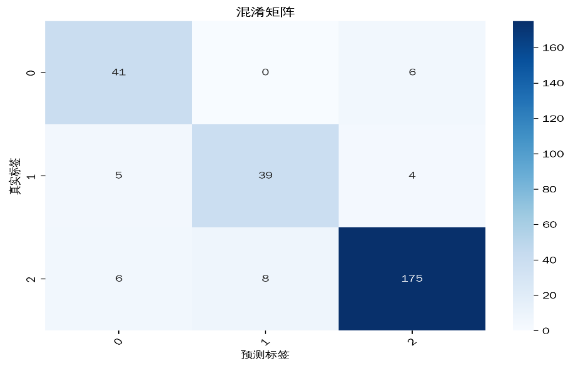
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 提示策略 | 准确率 | 精确率 | 召回率 | F1 分数 | AUC 值 |
| 零样本提示 | 0.704225352112676 | 0.6560363530055422 | 0.704225352112676 | 0.6784663247048188 | 0.6709964855486846 |
| 90 样本提示 | 0.897887323943662 | 0.9002509272025304 | 0.897887323943662 | 0.8986319612644122 | 0.9041985643947756 |
| 分解提示 | 0.9190140845070423 | 0.9191754537248555 | 0.9190140845070423 | 0.9188775418205605 | 0.9163083064019824 |

为了更直观地展示不同提示策略下模型的分类性能，我们绘制了混淆矩阵图和 ROC 曲线（ROC 曲线则通过展示模型在不同阈值下的真正率和假正率之间的关系，评估模型的分类性能 。曲线越靠近左上角，说明模型的分类性能越好 ） 。混淆矩阵图以矩阵的形式展示了模型预测结果与真实标签之间的关系，能够清晰地反映出模型在不同类别上的分类准确性和误分类情况 。

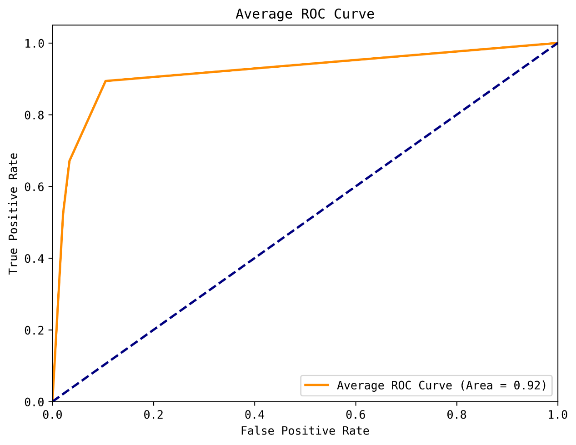
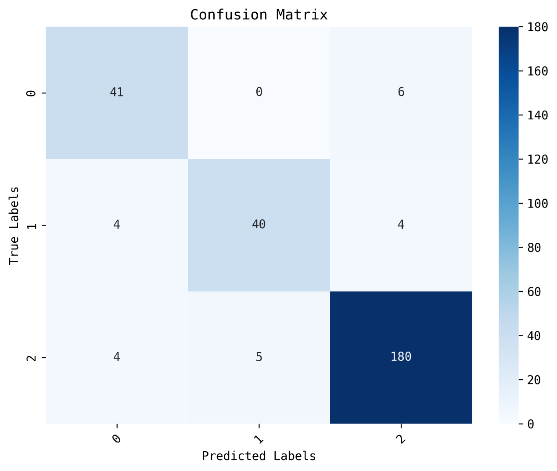
在零样本提示的混淆矩阵中，我们可以看到，模型在区分正常、睡眠呼吸暂停和失眠这三种睡眠障碍类型时，存在较多的误分类情况 。正常样本被误判为睡眠呼吸暂停或失眠的数量较多，睡眠呼吸暂停和失眠样本之间也存在一定程度的误判 。



在 90 样本提示的混淆矩阵中，误分类情况有了明显改善 。正常样本和睡眠呼吸暂停样本的正确分类数量增加，误判情况减少 。但在失眠样本的分类上，仍存在一定的误判情况。



分解提示的混淆矩阵显示，模型在各个类别上的分类准确性都有了显著提高 。正常样本、睡眠呼吸暂停样本和失眠样本的正确分类数量都达到了较高水平，误分类情况最少 。这体现出分解提示策略在睡眠障碍分类任务中的有效性和优越性 。



从绘制的 ROC 曲线可以看出，分解提示策略下的曲线最靠近左上角，AUC 值最大，表明其分类性能最佳 。90 样本提示策略的 ROC 曲线次之，零样本提示策略的 ROC 曲线最远离左上角，AUC 值最小，分类性能最差 。

### 5.4 实验结果分析

* 零样本提示在睡眠障碍分类任务中表现欠佳，主要原因在于其对数据利用的局限性和模型学习能力的不足 。零样本提示仅依据简单的预定义规则进行分类，缺乏对训练数据的深入学习 。在睡眠障碍分类任务中，睡眠障碍的判定受到多种因素的综合影响。这些因素相互交织，形成了复杂的非线性关系 。零样本提示无法从训练数据中学习到这些复杂的特征与分类结果之间的关系，难以捕捉数据中的潜在模式 ，导致分类性能欠佳 。
* 90 样本提示相比零样本提示，分类性能显著提升。大语言模型参考训练集数据模式、特征与分类结果的对应关系，能学习到不同特征组合和睡眠障碍类型的潜在联系，像分析训练集中性别、年龄、职业等因素与睡眠障碍类型的关系，可更好理解这些因素在分类中的作用。并且用逻辑回归模型分类，通过对类别型特征编码、数值型特征标准化等预处理提升数据质量，优化学习效果，在面对常见睡眠障碍类型时判断更准确。但逻辑回归模型是线性的，面对特征间复杂非线性关系的数据，分类能力有限，处理睡眠障碍与多种因素的复杂非线性关系时，可能无法准确捕捉，导致分类错误。
* 分解提示策略在睡眠障碍分类任务中取得了最佳效果，这主要归功于其对任务的有效拆解和对多种分类器的优化选择 。分解提示将任务进行分解，对逻辑回归、决策树、随机森林、梯度提升、支持向量机和 K 近邻等多种分类器进行尝试，并实施参数调优 。通过对不同分类器的比较与筛选，确定支持向量机（SVM）为最适配该数据集的模型 。支持向量机能够在高维空间中寻得最优分类超平面，有效处理复杂的非线性关系 。在睡眠障碍分类问题中，许多特征之间存在复杂的非线性联系，例如睡眠时长与睡眠质量、压力水平与日常活动量等 。SVM 能够精准捕捉这些关系，进而获得最佳分类性能 。在判断失眠与压力水平、生活习惯等因素的关系时，SVM 可以通过寻找最优分类超平面，准确地将失眠样本与其他样本区分开来 。分解提示对任务的分解使得模型能够更深入地理解任务要求，逐步完成数据处理、模型训练和评估等环节，提高了模型的分类准确性和可靠性 。

通过对不同提示策略的实验结果进行分析与对比，我们可以看出，在睡眠障碍分类任务中，充分利用训练数据、提升模型学习能力以及合理分解任务和选择分类器是提高模型性能的关键因素 。分解提示策略在处理复杂的睡眠障碍分类问题时具有明显的优势，为睡眠障碍分类提供了更有效的方法和思路 。

## 六、研究结论与展望

### 6.1 研究结论

在本次睡眠障碍分类研究中，我们深入探索了大语言模型在睡眠健康与生活方式数据集上的应用，通过精心设计的实验方案和严谨的分析过程，取得了一系列成果。大语言模型在睡眠障碍自动分类任务中展现出了显著的潜力。不同的提示策略对大语言模型的性能产生了关键影响：

1.零样本提示作为一种简单直接的提示策略，在睡眠障碍分类任务中表现相对较弱 。这表明，在缺乏具体示例和深入学习的情况下，大语言模型仅凭预训练知识和简单规则，难以准确应对睡眠障碍分类的复杂性 。

2.90 样本提示策略通过引入一定数量的样本数据，为大语言模型提供了更丰富的学习信息 。与零样本提示相比，其的分类性能有了显著提升 。这充分说明，利用训练数据中的模式和对应关系，能够增强大语言模型对睡眠障碍分类任务的理解和执行能力 。然而，90 样本提示策略仍存在一定的局限性 。由于其使用的逻辑回归模型具有线性特性，对于特征间存在复杂非线性关系的数据，难以准确捕捉和处理这些关系，从而限制了模型的分类能力 。

3.分解提示策略在睡眠障碍分类任务中取得了最为优异的成绩 。对多种分类器进行了全面的尝试和深入的参数调优 。经过评估和比较，合理分解任务和选择合适的分类器，表明其能够能够充分发挥大语言模型的优势，提高睡眠障碍分类的准确性和可靠性 。

在利用大语言模型进行睡眠障碍分类研究中，数据与任务理解对模型性能影响重大，直接关乎分类准确性。提供详细示例、参考信息和合理分解任务，有助于提升模型分类的准确性。因此，未来研究应着重深度剖析数据和任务，优化提示策略，以此增强模型性能。同时，在分解提示过程中，根据任务特性选择适配模型并加以优化是提升分类性能的核心要点。由于不同分类器适用场景各异，面对复杂的睡眠障碍分类任务时，需综合权衡多种因素，精准挑选最佳分类器并细致调整参数，从而实现最优分类效果。

### 6.3 研究局限及未来方向

本研究在睡眠障碍分类中虽然取得了一定的成果，但在数据、模型、实验设计、提示工程等方面仍存在局限性。如数据的多样性不足[15]、分类器模型的选择没有利用最新的深度学习强大的特征学习能力、对比实验不够充分和没有外部验证、优化提示策略不足。未来的研究可以从以上多个方向展开，以进一步提升大语言模型在睡眠障碍自动分类任务中的性能和应用价值 。

* 扩充数据集是当务之急，需要广泛收集各类睡眠障碍的样本，特别是罕见或特殊类型的睡眠障碍数据。同时，改善数据分布不均衡问题也是关键，可利用数据重采样技术，如过采样（SMOTE 算法）和欠采样[16]，调整数据集中不同类别样本的比例，让模型在训练过程中能更均衡地学习各类睡眠障碍的特征从而提高模型对特定群体的泛化能力。
* 在分类器模型改进方面，探索更先进的架构和特征工程方法至关重要 。深度学习模型具有强大的特征学习能力[17]，如卷积神经网络（CNN）[18]、循环神经网络（RNN）[19]及其变体长短期记忆网络（LSTM）[20]、门控循环单元（GRU）[21]等 。这些模型在处理时间序列数据和复杂数据结构方面具有优势，能够自动学习这些数据中的特征和模式，更好地挖掘睡眠障碍与难以量化因素之间的潜在关系。
* 在实验设计完善方面，应开展更全面的对比实验和外部验证 。全面对比不同的模型架构（包括LLM和分类模型）、提示策略等因素对分类性能的影响，能够帮助我们深入了解各种因素的作用机制，找到最适合睡眠障碍分类任务的方法和参数设置 。通过在多个不同来源的外部数据集上进行测试，可以评估模型在不同数据分布和场景下的泛化能力，避免模型出现过拟合问题。
* 在提示策略优化方面，可以探索使用更复杂的提示方式，如链式思维提示（Chain of Thought Prompting[22]）、情境提示（Contextual Prompting）[23]等，引导模型进行更深入的推理和分析。链式思维提示可以让模型在回答问题时，逐步展示其推理过程，提高模型的可解释性[24]。情境提示可以根据任务的具体情境，为模型提供更多的背景信息和上下文，帮助模型更好地理解任务要求，从而更准确地进行睡眠障碍分类 。

七、致谢

在完成这项研究的过程中，我得到了许多个人和机构的帮助，在此，我要向他们表达我最诚挚的感谢。

我要衷心感谢我的导师。在整个研究过程中，导师凭借其深厚的学术造诣和丰富的经验，为我指明了研究方向。当我在研究中遇到困惑和难题时，导师总是耐心地倾听我的问题，给予我宝贵的建议和指导。在数据集的分析和提示策略的设计阶段，导师的专业见解让我对研究有了更深入的理解，避免了许多弯路。导师严谨的治学态度和对科研的执着追求，也一直激励着我不断前进，是我学术道路上的榜样。

我同样要感谢数据提供方。他们为我们提供了宝贵的睡眠健康与生活方式数据集，这是整个研究的基础。。

在未来的学术道路上，我将继续努力，不断探索，为睡眠医学领域的发展贡献自己的力量。

八参考文献

[1] Ibomoiye Domor Mienye, N., Jere, N. (Year Unknown). Survey of Decision Trees: Concepts, Algorithms, and Applications [Literature Type]. In IEEE Xplore . <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10562290>  
[2] Kumari, A., Akhtar, M., Shah, R., et al. Support matrix machine: A review[EB/OL]. <https://arxiv.org/abs/2310.19717>.  
[3] Curth, A., Jeffares, A., van der Schaar, M. Why do Random Forests Work? Understanding Tree Ensembles as Self-Regularizing Adaptive Smoothers[EB/OL]. <https://arxiv.org/abs/2402.01502>.  
[4] Kim, Y., Xu, X., McDuff, D., et al. Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data[EB/OL].[Submitted on 12 Jan 2024 (v1), last revised 27 Apr 2024 (this version, v2)].  
[5] Katz, D. M., Bommarito, M. J., Gao, S., et al. GPT-4 passes the bar exam[J]. Philosophical Transactions of the Royal Society A, 2024, 382(2270):20230254.  
[6] Nori, H., Lee, Y. T., Zhang, S., et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine[EB/OL]. arXiv preprint arXiv:2311.16452, 2023.  
[7] Saab, K., Tu, T., Weng, W.-H., et al. Capabilities of Gemini models in medicine[EB/OL]. arXiv preprint arXiv:2404.18416, 2024.  
[8] Singhal, K., Tu, T., Gottweis, J., et al. Towards expert-level medical question answering with large language models[EB/OL]. arXiv preprint arXiv:2305.09617, 2023.  
[9] McDuff, D., Schaekermann, M., Tu, T., et al. Towards accurate differential diagnosis with large language models[EB/OL]. arXiv preprint arXiv:2312.00164, 2023.  
[10] Sleep Health and Lifestyle Dataset[EB/OL]. <https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset>.  
[11] Wang, G., Zhao, W., Han, J., et al. MedFound: The First Medical Large Language Model Passing the Physician Qualification Examination [J]. Journal of Artificial Intelligence, 2024, 5(1): 1 - 12.  
[12] McDuff, D., Xu, X., Kim, Y., et al. Personal Health Large Language Model (PH-LLM): leveraging large language models for personalized health insights [EB/OL]. arXiv preprint arXiv:2311.17133, 2023.  
[13] Zhang, Y., Maziarka, P., Klicpera, J., et al. DiffSBDD: Equivariant diffusion for structure-based drug design [EB/OL]. arXiv preprint arXiv:2403.14338, 2024.

[4]Vaswani, A., Shazeer, N., Parmar, N., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. In Advances in Neural Information Processing Systems 30 (NIPS 2017).

[15] Liu, X., Uchiyama, M., Okawa, M., et al. Prevalence and correlates of insomnia in the Japanese general population: Results from the Japan epidemiological sleep study[J]. Sleep, 2000, 23(4):497 - 506.  
[16] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. SMOTE: Synthetic Minority Over - sampling Technique[J]. J Artif Intell Res, 2002, 16:321 - 357.  
[17] LeCun, Y., Bengio, Y., Hinton, G. Deep Learning[J]. Nature, 2015, 521(7553):436 - 444.  
[18] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. Gradient - Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278 - 2324.  
[19] Graves, A., Mohamed, A. - R., Hinton, G. Speech Recognition with Deep Recurrent Neural Networks[C]. In Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645 - 6649), 2013.  
[20] Hochreiter, S., Schmidhuber, J. Long Short - Term Memory[J]. Neural Comput, 1997, 9(8):1735 - 1780.  
[21] Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y. On the properties of neural machine translation: Encoder–decoder approaches[EB/OL]. arXiv preprint arXiv:1409.1259, 2014.  
[22] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichtertz, R., Joshi, S., Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[EB/OL]. arXiv preprint arXiv:2201.11903, 2022.  
[23] Zhou, X., Huang, M., Wang, H., Zhang, Z. Contextual Prompting for Few - Shot Text Classification[C]. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 9312 - 9327), 2022.  
[24] Ribeiro, M. T., Singh, S., Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier[C]. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1135 - 1144), 2016.