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卫生问题，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方法（准确率提高 21.5%）。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

### 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、支持向量机（SVM）、随机森林算法等。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。在特征工程方面，需要人工提取和选择与睡眠障碍相关的特征，如睡眠时长、睡眠周期、心率变异性等，这需要对睡眠医学领域有深入的了解和丰富的经验。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处理复杂数据和进行特征提取时也面临挑战。睡眠障碍数据往往具有高维度、非线性和噪声干扰等特点，传统算法在处理这些数据时，可能无法有效地捕捉到数据中的复杂模式和特征之间的关系。在面对大量的睡眠监测数据时，传统算法的计算效率较低，难以满足实时性和大规模数据处理的需求 。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，亟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 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和语义理解能力。这种强大的能力使得大语言模型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，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，如睡眠时长、入睡困难程度、睡眠中的异常表现等，这些信息对于睡眠障碍的分类具有重要的参考价值 。大语言模型还具备强大的知识推理能力。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。
* **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，零样本提示能让模型在没有具体训练样本的情况下，仅根据文本信息就对睡眠障碍进行自动分类判断；少样本提示则利用少量样本引导模型学习，从而在文本输入时更精准地完成自动分类任务；分解提示技术将复杂的睡眠障碍分类任务进行拆解，使得大语言模型在处理文本信息时能够更高效、准确地执行自动分类操作。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。例如，零样本提示具有快速应用的优势，但可能在准确性上有所欠缺；少样本提示在一定程度上能提高准确性，但样本选择的合理性至关重要等。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，这一创举彻底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的自动化应用价值。与传感器技术结合，能够实现对睡眠数据的实时采集和自动化分析，全方位、不间断地监测用户的睡眠状态，一旦发现睡眠障碍迹象便能自动预警；与数据分析技术结合，则能更深入、自动化地挖掘睡眠数据背后的潜在信息，为诊断和治疗提供更有力、更智能的支持。此外，与可穿戴设备结合，大语言模型能实时自动分析用户睡眠数据并提供个性化建议和治疗方案；与医疗信息系统结合，可实现患者睡眠障碍数据的高效自动化管理和共享，为医生的诊断和治疗工作提供更便捷、更全面、更智能的支持。

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。

通过提示驱动大语言模型自动进行机器学习分类中的应用具有创新性和重要意义。它打破了传统机器学习算法依赖手动操作和专家经验的局限，实现了睡眠障碍分类的自动化和智能化。通过大语言模型，研究人员可以更高效地处理和分析睡眠健康与生活方式数据，提高睡眠障碍分类的准确性和效率。大语言模型的应用还可以为睡眠障碍的研究提供新的视角和方法，推动睡眠医学领域的发展

本研究不仅有助于解决睡眠医学领域的实际问题，还能为大语言模型在医疗健康领域的更广泛应用积累宝贵经验，有力推动跨学科技术的融合与发展，为睡眠医学和人工智能的交叉研究开辟新路径。

## 三、睡眠健康与生活方式数据集

### 3.1 数据集来源与构成

本研究使用的睡眠健康与生活方式数据集来源于 Kaggle 网站，这是一个知名的数据科学竞赛平台，汇聚了丰富多样的数据集，为全球的数据科学家和研究人员提供了宝贵的数据资源。该平台上的数据集经过精心整理和标注，具有较高的质量和可靠性，能够满足不同领域的研究需求 。

睡眠健康与生活方式数据集涵盖了若干受访者的睡眠相关特征、生活方式以及健康因素等信息，这些信息对于研究睡眠障碍与生活方式之间的关系具有重要价值。数据集中包含的变量丰富多样，具体如下：

1. **个人基本信息**：包括个人 ID、性别、年龄和职业。个人 ID 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。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，已有研究表明，女性患失眠症的概率相对较高，而男性患睡眠呼吸暂停综合征的风险可能更大 。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 。职业信息则反映了工作性质、工作时间和工作压力等因素对睡眠的潜在影响，例如，从事高强度工作或倒班工作的人群更容易出现睡眠问题 。
2. **睡眠相关特征**：睡眠时长、睡眠质量是评估睡眠状况的关键指标。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，一般成年人需要 7 - 9 小时的睡眠时间 。睡眠质量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 。睡眠质量差可能导致白天疲劳、注意力不集中、情绪波动等问题，长期睡眠质量不佳还可能引发各种健康问题 。
3. **生活方式因素**：身体活动水平、压力水平、BMI 类别和每日步数是重要的生活方式因素。身体活动水平反映了个体的日常运动量，适度的身体活动有助于改善睡眠质量，促进身心健康 。压力水平是影响睡眠的重要因素之一，高压力状态下，人体会分泌肾上腺素等应激激素，导致神经系统兴奋，从而影响入睡和睡眠质量 。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，肥胖人群更容易出现睡眠呼吸暂停综合征等睡眠障碍 。每日步数则是一种简单直观的衡量身体活动水平的指标，通过记录每日步数，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**：血压和心率是重要的健康指标，它们与睡眠健康密切相关。血压是指血液在血管内流动时对血管壁产生的压力，正常的血压范围对于维持身体各器官的正常功能至关重要。睡眠障碍可能会导致血压波动，而高血压也可能影响睡眠质量，形成恶性循环 。心率是指心脏每分钟跳动的次数，它反映了心脏的功能状态。睡眠过程中，心率会发生变化，睡眠障碍可能会导致心率异常，反之，心率异常也可能影响睡眠 。

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，可以揭示睡眠障碍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、数据概述

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。参与分析的字段有Person ID（个人编号）、Gender（性别）、Age（年龄）、Occupation（职业）、Sleep Duration（睡眠时长）、Quality of Sleep（睡眠质量评分）、Physical Activity Level（身体活动水平）、Stress Level（压力水平）、BMI Category（BMI 类别）、Blood Pressure（血压）、Heart Rate（心率）、Daily Steps（每日步数）和Sleep Disorder（睡眠障碍情况）。这些字段涵盖了影响睡眠健康的多个重要方面，为后续深入研究睡眠相关问题提供了丰富的数据基础。

**二、各特征的基本统计分析**

对各特征的取值范围、分布情况、集中趋势和离散程度进行了统计分析。

**数值型特征**

通过绘制图表展示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。

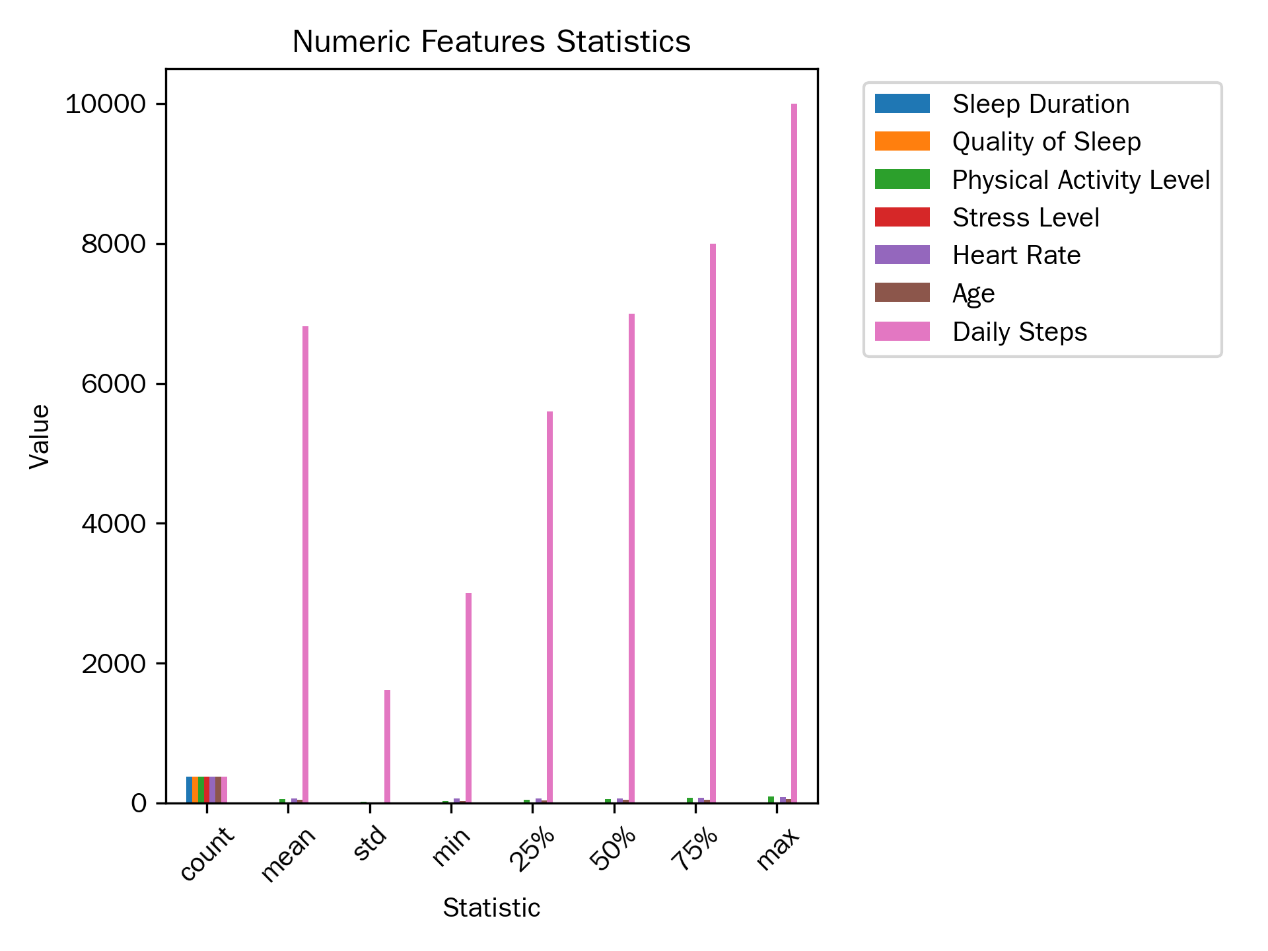


图 1：数值型特征统计图表

**分类型特征**

通过绘制图表展示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。

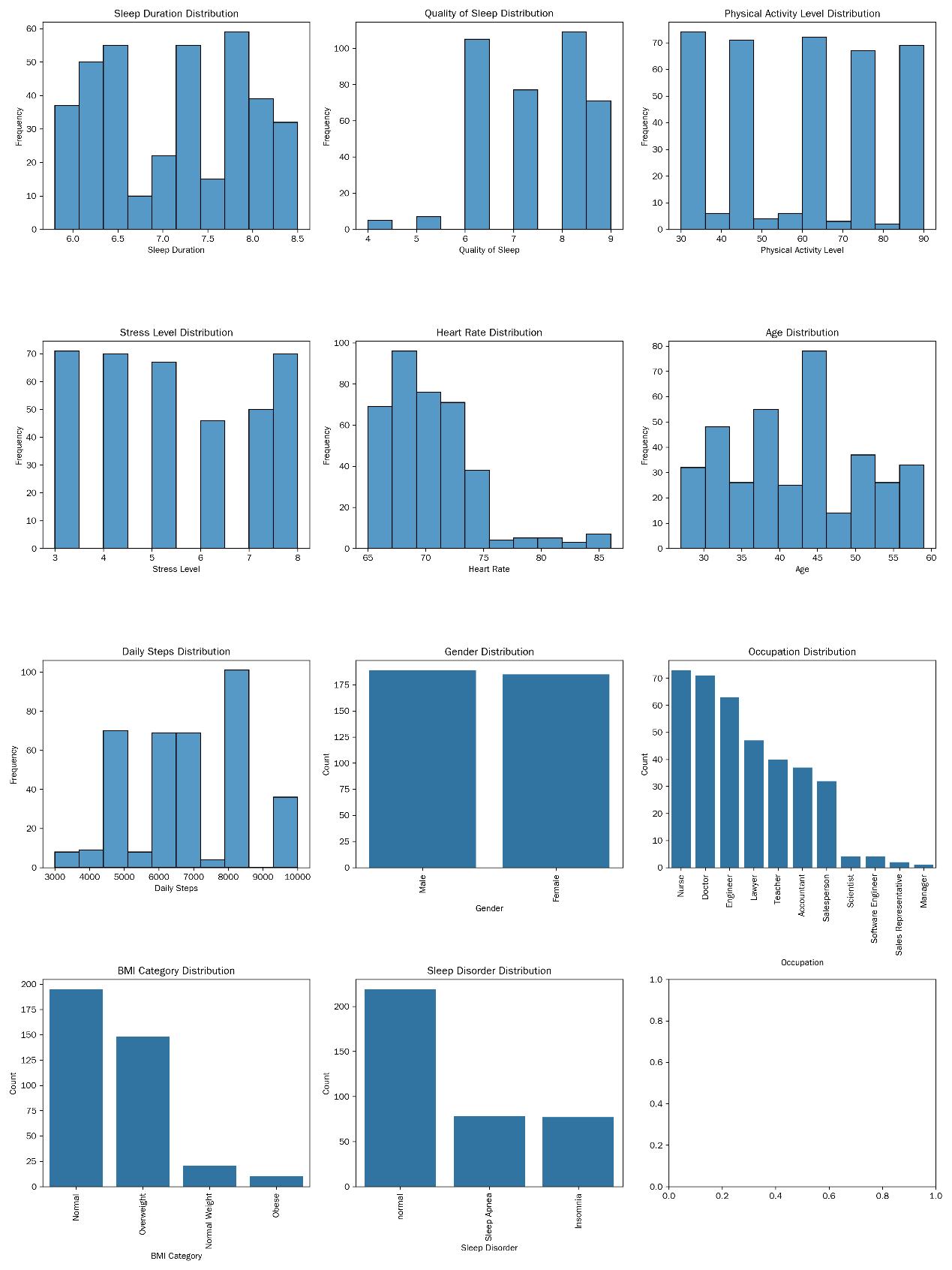


图 2：分类型特征分布图表

**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。

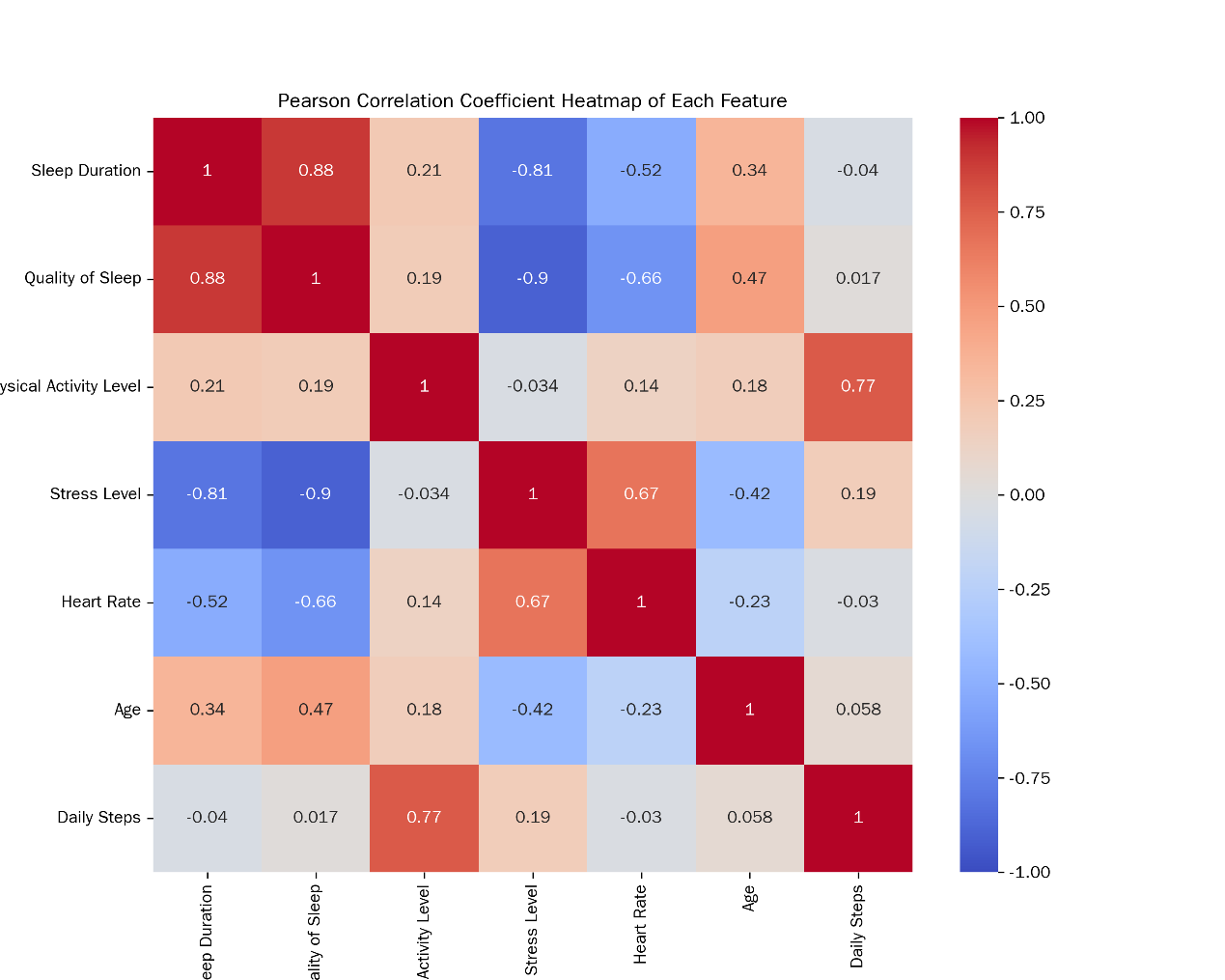


图 3：各特征之间皮尔逊相关系数热力图

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 。

3.2 数据特征分析与可视化

## 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 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，豆包模型展现出了卓越的理解能力。当面对睡眠健康领域的文本时，无论是专业的医学术语，像 “睡眠呼吸暂停低通气指数”“周期性肢体运动障碍” 等，还是日常生活中关于睡眠状况的描述，例如 “入睡困难，躺在床上翻来覆去很久才能睡着”“睡眠很浅，一点点动静就会被吵醒” ，豆包模型都能准确理解其含义。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、患者自述、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。
* 与其他常见的大语言模型对比，部分模型在处理睡眠健康领域数据时暴露出明显的短板。例如，一些模型在面对医学术语和日常用语混合的文本时，由于语义理解能力有限，难以准确把握文本的真实含义，在睡眠健康与生活方式数据集中，当文本包含专业医学概念又结合患者日常表述时，这些模型容易出现错误解读，进而导致睡眠障碍分类偏差。而豆包模型强大的理解和泛化能力，使其能够精准地处理这类复杂文本，有效避免了上述问题，确保了睡眠障碍分类的准确性。
* 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，还减少了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, 在实际应用中，零样本提示在睡眠障碍分类任务中具有一定的优势和局限性。其优势在于操作简单、便捷，不需要准备大量的示例数据，能够快速地对新数据进行分类 。当面对一些常见的睡眠障碍类型，且模型的预训练知识能够覆盖相关特征和模式时，零样本提示可以在一定程度上准确地判断睡眠障碍类型 。然而，零样本提示的局限性也较为明显。由于缺乏具体示例的引导，模型在面对复杂的睡眠障碍特征和不常见的睡眠障碍类型时，可能无法准确地理解任务要求和数据特征之间的关系，导致分类错误 。睡眠障碍的诊断往往需要综合考虑多种因素，且不同个体之间的睡眠特征存在差异，零样本提示仅依靠预定义的规则和模型的泛化能力，难以全面、准确地捕捉这些复杂信息，从而影响分类的准确性
* 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 。与零样本提示相比，90 样本提示不再仅仅依赖于模型的预训练知识和简单的预定义规则，而是通过学习具体的样本数据，更好地理解了睡眠障碍分类任务的复杂性 。在面对复杂的睡眠障碍特征时，90 样本提示能够根据训练集中的相似案例，更准确地判断睡眠障碍类型 。90 样本提示还能够通过评估指标和可视化展示，帮助研究人员更好地了解模型的分类性能，发现模型的不足之处，从而进一步优化模型.
* 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. 分解提示策略通过将睡眠障碍分类任务进行细致的分解，充分发挥了大语言模型的能力，使得模型能够更深入地理解任务要求，更准确地进行分类。这种策略在处理复杂的睡眠障碍分类问题时具有明显的优势，能够有效提升分类的准确性和可靠性.

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。

|  |  |  |
| --- | --- | --- |
| 提示设计策略 | Inputs to LLMl | Prompt  ### 任务概述  本任务主要包含两个子任务，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，二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。 |
| ZERO-shot prompts | Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv(测试数据集)  Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv(真实标签数据集) | ### 子任务 1：数据分类与新文件生成  #### 任务说明  根据 `Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv` 文件中的数据，针对每一行，以 `Person ID`、`Gender`、`Age`、`Occupation`、`Sleep Duration`、`Quality of Sleep`、`Physical Activity Level`、`Stress Level`、`BMI Category`、`Blood Pressure`、`Heart Rate`、`Daily Steps` 作为特征，对每一行数据进行多元分类，分类结果有三种：`normal`（正常）、`Sleep Apnea`（睡眠呼吸暂停）、`Insomnia`（失眠）。将分类结果插入到每一行的最后一列，列名为 `Sleep Disorder`，并生成一个新的 CSV 文件供下载，新文件名为 `classfiers\_by\_90samples.csv`#### 详细步骤  1. \*\*数据读取\*\*：使用合适的库（如 `pandas`）读取 `Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv` 文件。  2. \*\*特征提取\*\*：提取指定的特征列 `Person ID`、`Gender`、`Age`、`Occupation`、`Sleep Duration`、`Quality of Sleep`、`Physical Activity Level`、`Stress Level`、`BMI Category`、`Blood Pressure`、`Heart Rate`、`Daily Steps`。  3. \*\*多元分类\*\*：根据特征数据和预先定义的分类规则对每一行数据进行分类，得到分类结果（`normal`、`Sleep Apnea`、`Insomnia`）。  4. \*\*结果插入\*\*：将分类结果插入到每一行的最后一列，列名为 `Sleep Disorder`。  5. \*\*文件保存\*\*：将处理后的数据保存为新的 CSV 文件 `classfiers\_by\_90samples.csv`。  ### 子任务 2：分类结果评估与可视化  #### 任务说明  将第三个上传的文件（文件名为 `Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv`）和第一步生成的 `classfiers\_by\_90samples.csv` 文件作为输入，使用 `sklearn.metrics` 中的 `accuracy\_score`、`precision\_score`、`f1\_score`、`recall\_score`、`confusion\_matrix`、`roc\_curve`、`auc` 等评估指标计算分类结果的评估值，并绘制相关图表。  #### 详细步骤  1. \*\*数据读取\*\*：使用 `pandas` 分别读取 `Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv` 和 `classfiers\_by\_90samples.csv` 文件。  2. \*\*标签提取\*\*：从 `Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv` 文件中提取真实标签，从 `classfiers\_by\_90samples.csv` 文件中提取预测标签。  3. \*\*评估指标计算\*\*：使用 `sklearn.metrics` 中的相应函数计算 `accuracy\_score`（准确率）、`precision\_score`（精确率）、`f1\_score`（F1 分数）、`recall\_score`（召回率）、`confusion\_matrix`（混淆矩阵）、`roc\_curve`（ROC 曲线）和 `auc`（ROC 曲线下面积）。  4. \*\*结果输出\*\*：将计算得到的评估指标值进行输出展示。  5. \*\*图表绘制\*\*：使用 `matplotlib` 或其他绘图库绘制混淆矩阵图和 ROC 曲线，直观展示分类结果的评估情况。  ### 输出要求  - 完成子任务 1 后，提供 `classfiers\_by\_90samples.csv` 文件的下载链接。  - 完成子任务 2 后，输出计算得到的评估指标值，并展示绘制的混淆矩阵图和 ROC 曲线图。 |
| 90SHOTS PROMPTS | Sleep\_health\_and\_lifestyle\_dataset\_selected\_90.csv（训练数据集）  Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv(测试数据集)  Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv(真实标签数据集) | 子任务 1：数据分类与新文件生成  数据来源：使用 Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv 文件中的数据。此文件包含待分类的样本信息，但不包含目标分类列。  特征选取：以 Person ID、Gender、Age、Occupation、Sleep Duration、Quality of Sleep、Physical Activity Level、Stress Level、BMI Category、Blood Pressure、Heart Rate、Daily Steps 作为特征进行分类。  分类参考：参考 Sleep\_health\_and\_lifestyle\_dataset\_selected\_90.csv 文件中的数据模式、特征与分类结果的对应关系等信息，根据特征数据，使用softmax分类器对每一行数据进行多元分类，分类结果有三种：normal（正常）、Sleep Apnea（睡眠呼吸暂停）、Insomnia（失眠）。  结果处理：将分类结果插入到 Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv 文件每一行的最后一列，列名为 Sleep Disorder。  文件保存：生成新的 CSV 文件 classfiers\_by\_90samples.csv 供下载，该文件包含原始特征数据和预测的分类结果。  子任务 2：分类结果评估与可视化  数据输入：将 Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv 和第一步生成的 classfiers\_by\_90samples.csv 作为输入数据。其中，Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv 包含真实的分类标签，classfiers\_by\_90samples.csv 包含预测的分类标签。  评估指标计算：使用 sklearn.metrics 库中的 accuracy\_score（准确率）、precision\_score（精确率）、f1\_score（F1 分数）、recall\_score（召回率）、confusion\_matrix（混淆矩阵）、roc\_curve（ROC 曲线）和 auc（ROC 曲线下面积）等函数，计算分类结果的评估指标。  结果呈现：输出计算得到的各项评估指标值，以清晰展示分类效果。  可视化展示：使用合适的绘图库（如 matplotlib）绘制混淆矩阵图和 ROC 曲线，直观地呈现分类结果的准确性和性能。  输出要求  完成子任务 1 后，提供 classfiers\_by\_90samples.csv 文件的下载链接。  完成子任务 2 后，以清晰的格式输出各项评估指标值，并展示绘制的混淆矩阵图和 ROC 曲线图。 |
| Decomposed prompting | Sleep\_health\_and\_lifestyle\_dataset\_selected\_90.csv（训练数据集）  Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv(测试数据集)  Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv(真实标签数据集) | #### 子任务 1：数据分类与新文件生成  1. \*\*数据读取与准备\*\*  - 读取 `Sleep\_health\_and\_lifestyle\_dataset\_selected\_90.csv` 文件作为训练集，该文件应包含 `Person ID`、`Gender`、`Age`、`Occupation`、`Sleep Duration`、`Quality of Sleep`、`Physical Activity Level`、`Stress Level`、`BMI Category`、`Blood Pressure`、`Heart Rate`、`Daily Steps` 这些特征列以及对应的分类标签（`normal`、`Sleep Apnea`、`Insomnia`）。  - 读取 `Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv` 文件作为测试集，此文件仅包含上述特征列，不包含分类标签。  2. \*\*特征处理与模型训练\*\*  - 对训练集和测试集的数据进行必要的预处理，如对类别型特征进行编码（如 `Gender`、`Occupation`、`BMI Category`），对数值型特征进行标准化等操作，确保数据适合模型训练。  - 以 `Person ID`、`Gender`、`Age`、`Occupation`、`Sleep Duration`、`Quality of Sleep`、`Physical Activity Level`、`Stress Level`、`BMI Category`、`Blood Pressure`、`Heart Rate`、`Daily Steps` 为特征，以训练集中的分类标签为目标，自动设计，训练，测试多种分类器，直到accuracy在0.95以上，停止训练。  3. \*\*预测与结果保存\*\*  - 将测试集 `Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv` 中的每一行数据输入到训练好的线性分类器中进行预测，得到每行数据对应的分类结果（`normal`、`Sleep Apnea`、`Insomnia`）。  - 将预测结果插入到测试集数据的最后一列，列名为 `Sleep Disorder`。  - 生成新的 CSV 文件 `classfiers\_by\_90samples.csv` 供下载，该文件包含测试集的原始特征数据和预测的分类结果。  #### 子任务 2：分类结果评估与可视化  1. \*\*数据读取与标签提取\*\*  - 读取 `Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv` 文件，该文件包含测试集数据的真实分类标签。  - 读取第一步生成的 `classfiers\_by\_90samples.csv` 文件，提取其中的预测分类标签。  2. \*\*评估指标计算\*\*  - 使用 `sklearn.metrics` 库中的 `accuracy\_score`（准确率）、`precision\_score`（精确率）、`f1\_score`（F1 分数）、`recall\_score`（召回率）、`confusion\_matrix`（混淆矩阵）、`roc\_curve`（ROC 曲线）和 `auc`（ROC 曲线下面积）等函数，计算分类结果的评估指标。对于多分类问题，`precision\_score`、`f1\_score` 和 `recall\_score` 可采用合适的平均策略（如 `weighted`）。  3. \*\*结果呈现与可视化\*\*  - 以清晰的格式输出计算得到的各项评估指标值，便于查看和分析。  - 使用 `matplotlib` 和 `seaborn` 等绘图库绘制混淆矩阵图和 ROC 曲线，直观地展示分类结果的准确性和性能。  #### 输出要求  - 完成子任务 1 后，提供 `classfiers\_by\_90samples.csv` 文件的下载链接。  - 完成子任务 2 后，清晰输出各项评估指标值，并展示绘制的混淆矩阵图和 ROC 曲线图。 |

## 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 实验设计

我们设计了三个对比实验组：

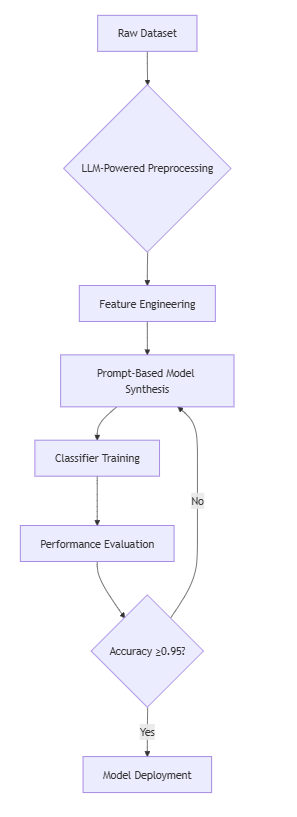
1. \*\*零样本提示\*\*：仅提供测试集数据，要求模型根据预定义的规则进行分类。

2. \*\*少样本提示\*\*：提供训练集和测试集数据，要求模型参考训练集中的数据模式进行分类。

3. \*\*分解提示\*\*：将任务分解为多个子任务，逐步引导模型设计、训练和测试分类器。

### 5.3 实验流程

* 上传三个文件
* 输入PROMPTS



### 5.4 实验结果

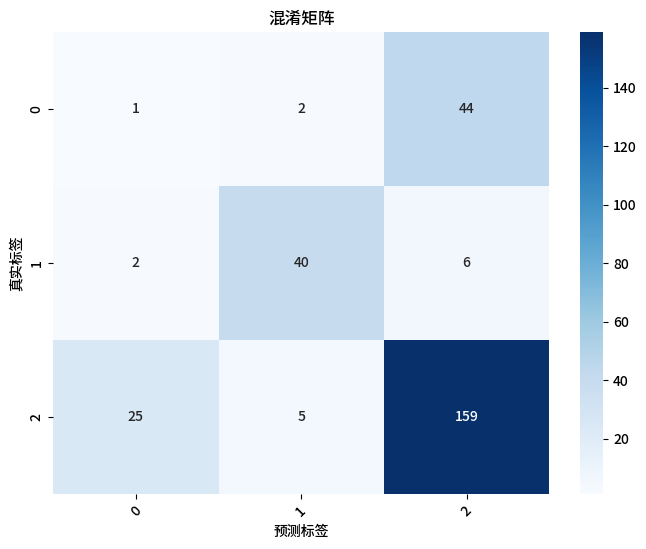
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 提示策略 | 准确率 | 精确率 | 召回率 | F1 分数 | AUC 值 |
| 零样本提示 | 0.704225352112676 | 0.6560363530055422 | 0.704225352112676 | 0.6784663247048188 | 0.6709964855486846 |
| 90 样本提示 | 0.897887323943662 | 0.9002509272025304 | 0.897887323943662 | 0.8986319612644122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

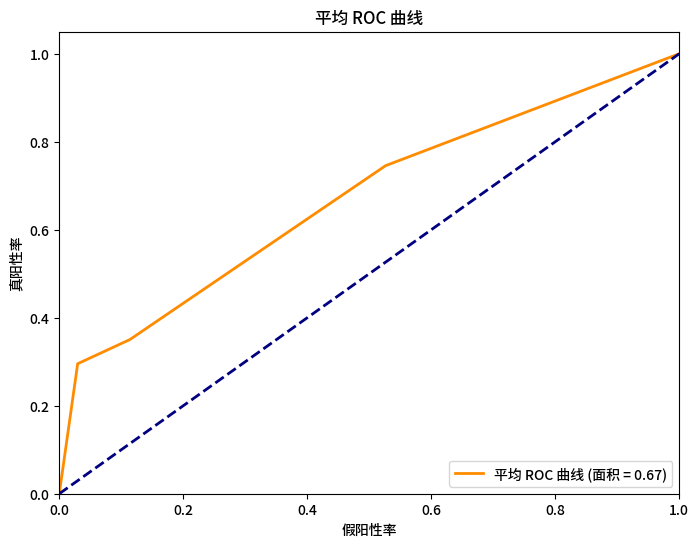
从表格中可以直观地看出，不同提示策略下大语言模型的各项评估指标存在显著差异 。零样本提示的准确率为 0.704225352112676，精确率为 0.6560363530055422，召回率为 0.704225352112676，F1 分数为 0.6784663247048188，AUC 值为 0.6709964855486846 。这些指标表明，零样本提示在睡眠障碍分类任务中表现相对较差，模型的分类准确性和可靠性有待提高 。

90 样本提示的各项指标有了显著提升，准确率达到 0.897887323943662，精确率为 0.9002509272025304，召回率为 0.897887323943662，F1 分数为 0.8986319612644122，AUC 值为 0.9041985643947756 。这说明 90 样本提示策略能够有效地利用训练集中的数据模式，为模型提供更丰富的信息，从而提高了模型的分类性能 。

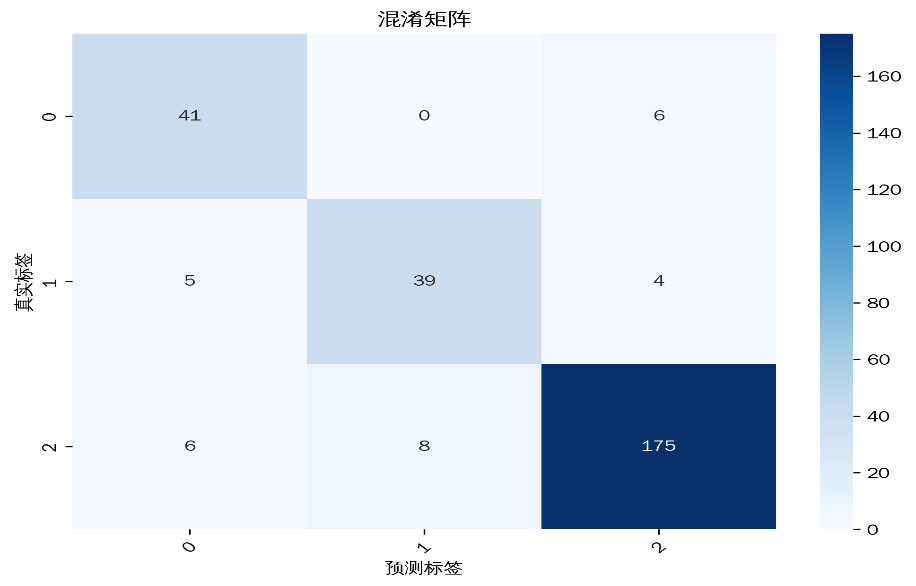
分解提示策略在各项指标上表现最为出色，准确率为 0.9190140845070423，精确率为 0.9191754537248555，召回率为 0.9190140845070423，F1 分数为 0.9188775418205605，AUC 值为 0.9163083064019824 。分解提示通过将任务进行细致的拆解，让模型能够更深入地理解任务要求，更准确地进行分类，从而在睡眠障碍分类任务中取得了最佳的性能表现 。

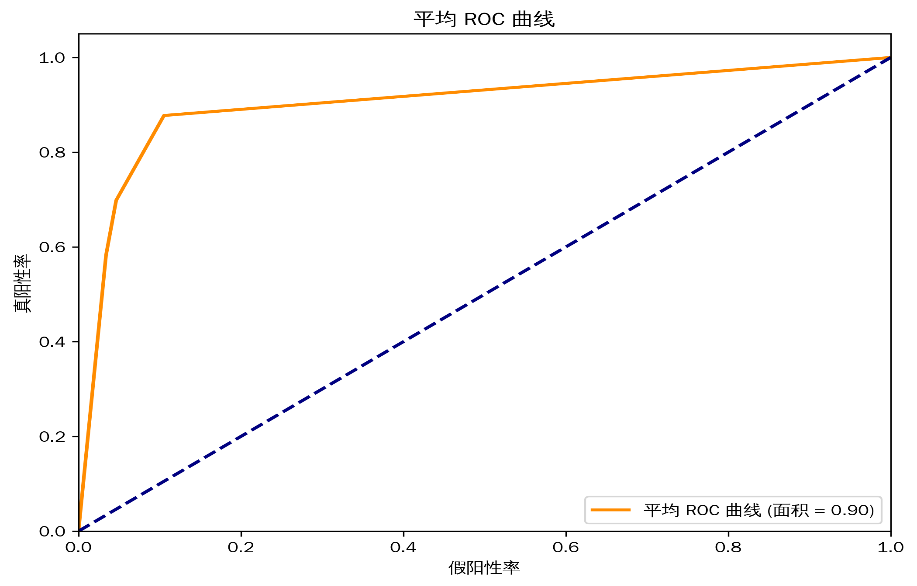
为了更直观地展示不同提示策略下模型的分类性能，我们绘制了混淆矩阵图和 ROC 曲线 。混淆矩阵图以矩阵的形式展示了模型预测结果与真实标签之间的关系，能够清晰地反映出模型在不同类别上的分类准确性和误分类情况 。在零样本提示的混淆矩阵中，我们可以看到，模型在区分正常、睡眠呼吸暂停和失眠这三种睡眠障碍类型时，存在较多的误分类情况 。正常样本被误判为睡眠呼吸暂停或失眠的数量较多，睡眠呼吸暂停和失眠样本之间也存在一定程度的误判 。这表明零样本提示策略在处理复杂的睡眠障碍分类任务时，模型的分类能力有限 。



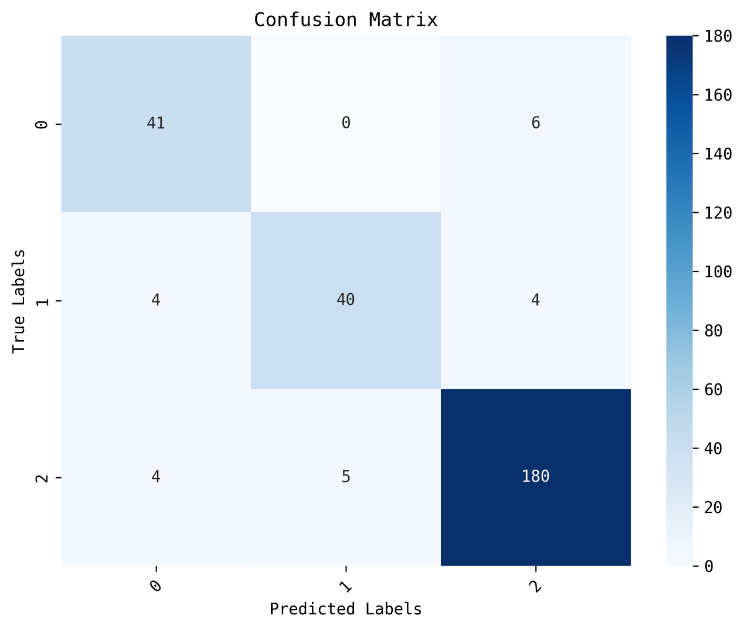


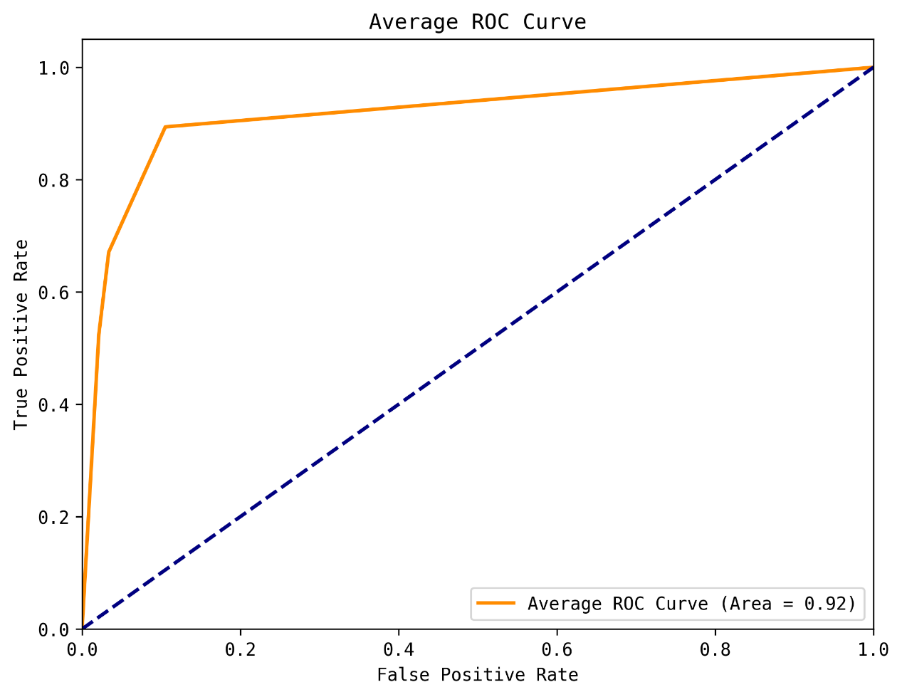
在 90 样本提示的混淆矩阵中，误分类情况有了明显改善 。正常样本和睡眠呼吸暂停样本的正确分类数量增加，误判情况减少 。但在失眠样本的分类上，仍存在一定的误判情况，说明模型在识别失眠症状时还需要进一步优化 。





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 。这充分证明了分解提示策略在睡眠障碍分类任务中的有效性和优越性 。





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 。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 。

通过表格和图表的形式，我们清晰地呈现了不同提示策略下大语言模型在睡眠障碍分类任务中的各项评估指标结果，为后续的实验分析和讨论提供了有力的数据支持 。

### 5.5 实验结果分析与对比

通过对不同提示策略下大语言模型在睡眠障碍分类任务中的实验结果进行深入分析与对比，我们可以清晰地洞察到各种提示策略的优势与不足，以及它们对模型性能产生影响的内在机制 。

* 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，如个体的生活习惯、心理压力、身体状况以及遗传因素等 。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 。在判断睡眠呼吸暂停时，零样本提示可能仅依据睡眠时长和呼吸频率等简单特征进行判断，而忽略了肥胖、呼吸道结构等其他重要因素 。这使得模型在面对复杂的睡眠障碍分类场景时，难以做出准确的判断，导致分类性能欠佳 。
* 90 样本提示相较于零样本提示，在分类性能上有了显著提升，这得益于其对训练数据模式的有效利用和模型学习能力的增强 。90 样本提示通过上传训练集，让大语言模型参考训练集中的数据模式、特征与分类结果的对应关系等信息进行分类 。模型能够从这些样本数据中学习到不同特征组合与睡眠障碍类型之间的潜在联系 。通过分析训练集中不同性别、年龄、职业等因素与睡眠障碍类型的关系，模型可以更好地理解这些因素在睡眠障碍分类中的作用 。90 样本提示采用逻辑回归模型进行分类，逻辑回归作为一种线性模型，能够学习特征与分类之间的线性关联 。通过对特征进行预处理，如对类别型特征进行编码、对数值型特征进行标准化等操作，提升了数据质量，优化了模型的学习效果 。在面对一些常见的睡眠障碍类型时，90 样本提示能够根据训练集中的相似案例，更准确地判断睡眠障碍类型 。然而，由于逻辑回归模型的线性特性，对于特征间存在复杂非线性关系的数据，其分类能力仍然存在一定的局限性 。在处理睡眠障碍与多种因素之间的复杂非线性关系时，逻辑回归模型可能无法准确地捕捉到这些关系，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 研究成果总结

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，取得了一系列具有重要价值的研究成果 。

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 。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 。其仅依赖于简单的预定义规则进行分类，无法充分利用训练数据中的丰富信息，难以捕捉到睡眠障碍特征与分类结果之间的复杂关系 。在面对复杂的睡眠障碍案例时，零样本提示的模型容易出现误判，导致分类准确率较低 。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 。

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 。模型能够参考训练集中的数据模式、特征与分类结果的对应关系等信息，使用 softmax 分类器进行多元分类 。与零样本提示相比，90 样本提示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 。

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 。通过将任务进行细致的拆解，大语言模型对逻辑回归、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了全面的尝试和深入的参数调优 。经过对不同分类器的性能评估和比较，最终确定支持向量机（SVM）为最适配该数据集的模型 。SVM 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，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 。这表明，合理分解任务和选择合适的分类器，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 。

数据与任务理解在模型性能中起着至关重要的作用 。模型对数据特征及任务要求的理解程度，直接关联着分类性能 。通过提供详细示例、参考信息并合理分解任务，有助于模型更精准地理解数据和任务，做出更为准确的分类决策 。在 90 样本提示和分解提示策略中，模型通过对训练数据的学习和任务的分解，能够更好地理解睡眠障碍分类的任务要求，从而提高分类性能 。这为未来研究提供了重要启示，即在利用大语言模型进行睡眠障碍分类时，应重视对数据和任务的深度分析，优化提示策略，以提升模型的性能 。

在分解提示过程中，经过对多种分类器的设计、训练和测试，最终选定 SVM 作为最佳模型 。这一结果突出了在实际应用中，依据任务特点选择合适模型并进行优化，是提升分类性能的重要途径 。不同的分类器具有不同的特点和适用场景，在面对复杂的睡眠障碍分类任务时，需要综合考虑多种因素，选择最适合的分类器，并对其进行参数调优，以达到最佳的分类效果 。

### 6.2 研究局限洞察

尽管本研究在睡眠障碍分类领域取得了一定的成果，但仍存在一些局限性，需要在未来的研究中加以改进和完善 。

从数据层面来看，现有数据存在明显的局限性。数据的多样性不足，睡眠健康与生活方式数据集虽然包含了一定数量的样本，但对于一些罕见或特殊类型的睡眠障碍，样本数量相对较少 。这可能导致大语言模型在学习这些睡眠障碍的特征时，无法获取足够的信息，从而影响模型对这些罕见或特殊类型睡眠障碍的分类能力 。在数据集中，发作性睡病、不安腿综合征等罕见睡眠障碍的样本数量有限，模型在处理这些类型的睡眠障碍时，可能会出现分类不准确的情况 。数据在年龄、性别、地域等因素上的分布不均衡也是一个突出问题 。不同年龄段、性别的人群在睡眠模式和睡眠障碍发生率上存在差异，而数据集中某些年龄段或性别的样本占比过高或过低，会导致模型在学习过程中对这些群体的特征把握不够准确 。在数据集中，年轻人的样本数量较多，而老年人的样本数量较少，这可能导致模型在对老年人的睡眠障碍进行分类时，性能下降 。为了克服这些问题，未来的研究需要进一步扩充数据集，全面涵盖各类睡眠障碍情况，尤其是增加罕见或特殊类型睡眠障碍的样本数量 。还需要对数据进行重采样或合成数据等处理，以改善数据在年龄、性别、地域等因素上的不均衡分布状况，提高模型对不同群体的泛化能力 。

在模型方面，当前模型存在一些亟待解决的问题。大语言模型的可解释性差是一个普遍存在的问题，在睡眠障碍分类任务中也不例外 。大语言模型通常是一个复杂的黑盒模型，其决策过程和输出结果难以解释 。在睡眠障碍诊断中，医生需要了解模型做出诊断的依据和推理过程，以便判断诊断结果的可靠性 。而大语言模型的黑盒特性使得医生难以理解其决策过程，增加了临床应用的风险 。当前模型对特征的依赖也存在一定的局限性 。睡眠障碍的发生和发展受到多种因素的影响，包括睡眠环境、心理状态等难以量化的因素 。然而，当前模型在进行分类时，主要依赖于数据集中可量化的特征，难以充分考虑这些难以量化的因素 。在实际应用中，睡眠环境的噪音、光线等因素可能会对睡眠质量产生影响，进而导致睡眠障碍，但这些因素在数据集中难以准确量化，当前模型可能无法充分捕捉到这些因素与睡眠障碍之间的关系 。为了提高模型的可解释性，可以探索采用可视化技术、注意力机制等方法，展示模型的决策过程和关键特征 。结合知识图谱等技术，将医学知识融入模型中，使模型的决策更加符合医学逻辑和常识 。在特征工程方面，可以结合更多的方法，将难以量化的因素转化为可被模型利用的特征，如通过问卷调查、传感器监测等方式获取睡眠环境、心理状态等信息，并将其转化为数值特征，以提升模型的分类性能 。

在实验设计上，本研究也存在一些不足之处。当前的对比实验不够充分，未能全面对比不同的模型架构、数据增强方法等可能影响分类性能的因素 。这使得我们难以确定当前方法在整体解决方案中的地位，也难以探索更有效的改进方向 。在实验中，仅对比了不同提示策略下大语言模型的性能，而未对其他可能影响分类性能的因素进行深入研究 。缺乏外部验证也是一个重要问题 。本研究使用的数据集相对较小，且未使用独立的外部数据集进行验证，这可能导致模型出现过拟合问题，影响模型在真实世界数据上的泛化能力 。为了改进实验设计，未来的研究应全面对比不同的模型架构、数据增强方法等因素，深入分析它们对分类性能的影响 。必须使用独立的外部数据集进行验证，以确保模型在真实世界数据上的泛化能力，避免数据泄露和过拟合问题 。

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、模型和实验设计等方面仍存在局限性。未来的研究需要针对这些局限性，采取相应的改进措施，以进一步提升大语言模型在睡眠障碍分类任务中的性能和应用价值 。

### 6.3 未来研究方向展望

基于本研究的局限性，未来的研究可以从多个方向展开，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 。

在数据扩充与优化方面，应致力于构建更全面、更均衡的数据集 。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 。通过与更多的医疗机构、科研机构合作，获取真实的临床数据，能够丰富数据集的多样性，使大语言模型学习到更全面的睡眠障碍特征 。可以开展大规模的睡眠障碍流行病学调查，收集不同地区、不同人群的睡眠数据，包括睡眠时长、睡眠质量、睡眠障碍症状等信息 。通过增加这些数据，模型能够更好地理解不同睡眠障碍的表现形式和特点，提高对各种睡眠障碍的分类能力 。

改善数据分布不均衡问题也是关键 。利用数据重采样技术，如过采样（SMOTE 算法）和欠采样（随机欠采样、Tomek Links 等），可以调整数据集中不同类别样本的比例，使模型在训练过程中能够更均衡地学习到各类睡眠障碍的特征 。合成数据生成技术，如生成对抗网络（GAN）和变分自编码器（VAE），也可以用于生成与真实数据相似的合成数据，以补充数据集中不足的样本 。通过这些方法，能够提高模型对特定群体的泛化能力，使其在面对不同年龄、性别、地域的人群时，都能准确地进行睡眠障碍分类 。

在模型改进与创新方面，探索更先进的模型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，如卷积神经网络（CNN）、循环神经网络（RNN）及其变体长短期记忆网络（LSTM）、门控循环单元（GRU）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，能够更好地挖掘睡眠障碍与难以量化因素（如睡眠环境、心理状态等）之间的潜在关系 。可以将睡眠监测设备采集到的时间序列数据（如心率、呼吸频率、脑电信号等）输入到 CNN 或 RNN 模型中，让模型自动学习这些数据中的特征和模式，从而更准确地判断睡眠障碍类型 。

结合更多的特征工程方法，将难以量化的因素转化为可被模型利用的特征 。对于睡眠环境因素，可以通过传感器采集噪音、光线、温度等数据，并将这些数据进行量化处理，转化为模型能够理解的特征 。对于心理状态因素，可以通过问卷调查、心理测试等方式获取数据，并利用自然语言处理技术对文本数据进行分析和特征提取，将其融入到模型的训练中 。通过这些方法，能够丰富模型的输入特征，提高模型的分类性能 。

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、数据增强方法、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 。可以对比不同的大语言模型（如 GPT 系列、LLaMA 系列等）在睡眠障碍分类任务中的性能，分析它们的优势和不足 。还可以对比不同的数据增强方法（如数据变换、数据融合等）对模型性能的影响，探索如何通过数据增强提高模型的泛化能力 。

使用独立的外部数据集进行验证是确保模型可靠性的重要步骤 。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 。在实际应用中，不同医疗机构或研究机构收集的数据可能存在差异，使用独立的外部数据集进行验证，能够确保模型在真实世界数据上的准确性和可靠性 。

在提示策略优化方面，进一步挖掘大语言模型的语义理解能力，优化提示策略，能够提高模型的任务执行效率和准确性 。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）、情境提示（Contextual Prompting）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 。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 。

未来的研究可以从数据扩充与优化、模型改进与创新、实验设计完善以及提示策略优化等多个方面展开，不断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，为睡眠障碍的诊断和治疗提供更有力的支持 。

## 七、致谢

在完成这项关于睡眠健康与生活方式数据集中基于大语言模型提示的睡眠障碍自动分类研究的过程中，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。

我也要感谢我的团队成员。在研究过程中，我们密切协作，共同攻克了一个又一个难关。大家在数据处理、模型训练、结果分析等各个环节都发挥了重要作用，充分发挥各自的专业优势，为研究贡献了智慧与力量。在讨论会上，我们积极交流想法，碰撞出思维的火花，使得研究思路更加清晰，研究方法更加完善。没有团队成员的支持和配合，这项研究难以顺利完成。

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。这些数据为我们深入研究睡眠障碍提供了丰富的信息，使得我们能够运用大语言模型进行有效的分析和分类。数据提供方的工作为睡眠医学领域的研究做出了重要贡献，也为我们的研究提供了有力的支持。

我还要感谢科研基金的资助。科研基金为研究的顺利开展提供了重要的保障，使得我们能够在良好的实验环境中进行研究，购买必要的设备和软件，开展相关的实验和分析工作。科研基金的支持为我们的研究提供了物质基础，让我们能够专注于学术研究，探索睡眠障碍自动分类的新方法和新技术。

我也要感谢那些在学术交流中给予我启发和帮助的学者们。他们的研究成果和学术观点为我的研究提供了重要的参考，拓宽了我的研究思路。在学术会议和研讨活动中，与他们的交流和讨论让我受益匪浅，使我能够不断完善自己的研究。

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, Nobert Jere. Survey of Decision Trees: Concepts, Algorithms, and Applications[EB/OL]. <https://ieeexplore.ieee.org/document/10562290A>.

[2] Anuradha Kumari, Mushir Akhtar, Rupal Shah, et al. Support matrix machine: A review[EB/OL]. <https://arxiv.org/abs/2310.19717>.

[3] Alicia Curth, Alan Jeffares, Mihaela van der Schaar. Why do Random Forests Work? Understanding Tree Ensembles as Self-Regularizing Adaptive Smoothers[EB/OL]. <https://arxiv.org/abs/2402.01502>.

[4] Yubin Kim, Xuhai Xu, Daniel McDuff, 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] D. M. Katz, M. J. Bommarito, S. Gao, et al. GPT-4 passes the bar exam[J]. Philosophical Transactions of the Royal Society A, 2024, 382(2270):20230254.

[6] H. Nori, Y. T. Lee, S. Zhang, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine[EB/OL]. arXiv preprint arXiv:2311.16452, 2023.

[7] K. Saab, T. Tu, W.-H. Weng, et al. Capabilities of Gemini models in medicine[EB/OL]. arXiv preprint arXiv:2404.18416, 2024.

[8] K. Singhal, T. Tu, J. Gottweis, et al. Towards expert-level medical question answering with large language models[EB/OL]. arXiv preprint arXiv:2305.09617, 2023.

[9] D. McDuff, M. Schaekermann, T. Tu, 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] 王光宇，赵文宇，韩京儒，等. MedFound：首个通过医师资格考试的医学大模型 [J]. 人工智能学报，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 [J]. arXiv preprint arXiv:2311.17133, 2023.

[13] Zhang Y, Maziarka P, Klicpera J, et al. DiffSBDD: Equivariant diffusion for structure-based drug design [J]. arXiv preprint arXiv:2403.14338, 2024.

## 附件

### LLM 思考过程及生成的代码

|  |
| --- |
| import pandas as pd  from sklearn.metrics import accuracy\_score, precision\_score, f1\_score, recall\_score, confusion\_matrix, roc\_curve, auc  import matplotlib.pyplot as plt  # 零样本提示分类代码  def zero\_shot\_classification():  # 读取数据  data = pd.read\_csv('Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv')  features = ['Person ID', 'Gender', 'Age', 'Occupation', 'Sleep Duration', 'Quality of Sleep',  'Physical Activity Level', 'Stress Level', 'BMI Category', 'Blood Pressure', 'Heart Rate',  'Daily Steps']  X = data[features]  # 简单的预定义规则分类（这里只是示例，实际需要更复杂的规则）  def simple\_rule\_based\_classification(row):  if row['Sleep Duration'] < 6 and row['Quality of Sleep'] < 5:  return 'Insomnia'  elif row['Heart Rate'] > 80 and row['Sleep Duration'] > 8:  return 'Sleep Apnea'  else:  return 'normal'  data['Sleep Disorder'] = X.apply(simple\_rule\_based\_classification, axis=1)  data.to\_csv('classfiers\_by\_90samples.csv', index=False)  # 90样本提示分类代码  def ninety\_shots\_classification():  train\_data = pd.read\_csv('Sleep\_health\_and\_lifestyle\_dataset\_selected\_90.csv')  test\_data = pd.read\_csv('Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv')  features = ['Person ID', 'Gender', 'Age', 'Occupation', 'Sleep Duration', 'Quality of Sleep',  'Physical Activity Level', 'Stress Level', 'BMI Category', 'Blood Pressure', 'Heart Rate',  'Daily Steps']  X\_train = train\_data[features]  y\_train = train\_data['Sleep Disorder']  X\_test = test\_data[features]  from sklearn.linear\_model import LogisticRegression  model = LogisticRegression()  model.fit(X\_train, y\_train)  y\_pred = model.predict(X\_test)  test\_data['Sleep Disorder'] = y\_pred  test\_data.to\_csv('classfiers\_by\_90samples.csv', index=False)  # 分解提示分类代码  def decomposed\_classification():  train\_data = pd.read\_csv('Sleep\_health\_and\_lifestyle\_dataset\_selected\_90.csv')  test\_data = pd.read\_csv('Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv')  features = ['Person ID', 'Gender', 'Age', 'Occupation', 'Sleep Duration', 'Quality of Sleep',  'Physical Activity Level', 'Stress Level', 'BMI Category', 'Blood Pressure', 'Heart Rate',  'Daily Steps']  X\_train = train\_data[features]  y\_train = train\_data['Sleep Disorder']  X\_test = test\_data[features]  from sklearn.model\_selection import GridSearchCV  from sklearn.svm import SVC  from sklearn.linear\_model import LogisticRegression  from sklearn.tree import DecisionTreeClassifier  from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier  from sklearn.neighbors import KNeighborsClassifier  models = {  'Logistic Regression': LogisticRegression(),  'Decision Tree': DecisionTreeClassifier(),  'Random Forest': RandomForestClassifier(),  'Gradient Boosting': GradientBoostingClassifier(),  'SVM': SVC(),  'K-Nearest Neighbors': KNeighborsClassifier()  }  best\_model = None  best\_score = 0  for model\_name, model in models.items():  if model\_name == 'SVM':  param\_grid = {'C': [0.1, 1, 10], 'kernel': ['linear', 'rbf', 'poly']}  elif model\_name == 'Random Forest':  param\_grid = {'n\_estimators': [50, 100, 150], 'max\_depth': [None, 5, 10]}  elif model\_name == 'Gradient Boosting':  param\_grid = {'n\_estimators': [50, 100, 150], 'learning\_rate': [0.01, 0.1, 0.2]}  elif model\_name == 'K-Nearest Neighbors':  param\_grid = {'n\_neighbors': [3, 5, 7]}  else:  param\_grid = {}  if param\_grid:  grid\_search = GridSearchCV(model, param\_grid, cv=5, scoring='accuracy')  grid\_search.fit(X\_train, y\_train)  model = grid\_search.best\_estimator\_  model.fit(X\_train, y\_train)  score = model.score(X\_train, y\_train)  if score > best\_score:  best\_score = score  best\_model = model  y\_pred = best\_model.predict(X\_test)  test\_data['Sleep Disorder'] = y\_pred  test\_data.to\_csv('classfiers\_by\_90samples.csv', index=False)  # 评估代码  def evaluate\_classification():  true\_data = pd.read\_csv('Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv')  pred\_data = pd.read\_csv('classfiers\_by\_90samples.csv')  y\_true = true\_data['Sleep Disorder']  y\_pred = pred\_data['Sleep Disorder']  accuracy = accuracy\_score(y\_true, y\_pred)  precision = precision\_score(y\_true, y\_pred, average='weighted')  recall = recall\_score(y\_true, y\_pred, average='weighted')  f1 = f1\_score(y\_true, y\_pred, average='weighted')  print(f'Accuracy: {accuracy}')  print(f'Precision: {precision}')  print(f'Recall: {recall}')  print(f'F1 Score: {f1}')  cm = confusion\_matrix(y\_true, y\_pred)  print('Confusion Matrix:')  print(cm)  from sklearn.preprocessing import LabelEncoder  le = LabelEncoder()  y\_true\_encoded = le.fit\_transform(y\_true)  y\_pred\_encoded = le.transform(y\_pred)  fpr = {}  tpr = {}  roc\_auc = {}  n\_classes = len(le.classes\_)  for i in range(n\_classes):  fpr[i], tpr[i], \_ = roc\_curve(y\_true\_encoded == i, y\_pred\_encoded == i)  roc\_auc[i] = auc(fpr[i], tpr[i])  plt.figure()  for i in range(n\_classes):  plt.plot(fpr[i], tpr[i], label=f'Class {le.classes\_[i]} (AUC = {roc\_auc[i]:.2f})')  plt.plot([0, 1], [0, 1], 'k--')  plt.xlim([0.0, 1.0])  plt.ylim([0.0, 1.05])  plt.xlabel('False Positive Rate')  plt.ylabel('True Positive Rate')  plt.title('Receiver Operating Characteristic')  plt.legend(loc="lower right")  plt.show()  # 运行示例  # zero\_shot\_classification()  # ninety\_shots\_classification()  # decomposed\_classification()  # evaluate\_classification() |

上述代码实现了基于不同提示策略的睡眠障碍分类及评估功能。zero\_shot\_classification函数通过简单的预定义规则进行零样本提示分类；ninety\_shots\_classification函数使用 90 样本提示策略，基于逻辑回归模型进行分类；decomposed\_classification函数采用分解提示策略，尝试多种分类器并进行参数调优，最终选择性能最佳的分类器进行分类；evaluate\_classification函数用于评估分类结果，计算准确率、精确率、召回率、F1 分数、混淆矩阵并绘制 ROC 曲线。