arXiv:2602.08254v1 [cs.AI] 9 Feb 2026

SynthAgent：用于真实患者模拟的多智能体大语言模型框架——肥胖伴心理健康共病案例研究

Arman Aghaee*[1]、Sepehr Asgarian†[1]和Jouhyun Jeon‡[1]

[1]*Klick Health*，多伦多，安大略省，加拿大

摘要

模拟高保真患者为研究复杂疾病提供了有力途径，同时有助于应对真实世界数据零散、存在偏差且受隐私限制等挑战。本研究介绍了SynthAgent，这是一种新型多智能体系统（MAS）框架，旨在对伴有抑郁症、焦虑症、社交恐惧症和暴食症等共病精神障碍的肥胖患者进行建模。SynthAgent整合了来自理赔数据、人群调查和以患者为中心的文献中的临床及医学证据，构建融入人格特质的个性化虚拟患者；这些人格特质会影响依从性、情绪调节和生活方式行为。通过自主智能体交互，该系统可在多种社会心理情境下模拟疾病进展、治疗反应和生活管理。对100多名生成患者的评估表明，在所提出的多智能体系统框架中，GPT-5和Claude 4.5 Sonnet作为核心引擎实现了最高保真度，表现优于Gemini 2.5 Pro和DeepSeek-R1。因此，SynthAgent提供了一个可扩展且保护隐私的框架，可用于探索医疗和心理领域中的患者历程、行为动态及决策过程。

1 引言

肥胖是一种复杂、多维的疾病，不仅涉及代谢功能障碍，还涉及行为、情绪和心理方面。与普通人群相比，肥胖者更容易患上抑郁症、焦虑症、社交恐惧症以及暴食症等饮食失调行为。近期研究表明，肥胖与抑郁症之间存在双向关系：在多个队列中，二者会使对方的发病风险增加约 1.5 倍 [1]。在临床环境中，心理健康共病的并存使肥胖治疗更加复杂。抑郁症状可能降低身体活动的动力 [2]，而焦虑症和情绪失调可能引发饮食失调行为 [3]。总体而言，心理健康方面的挑战与较差的治疗依从性、对标准疗法的反应减弱以及较低的肥胖者生活质量 [4]相关。因此，有效的肥胖症护理模式需要纳入心理、行为和情境维度，而不是只关注体重和代谢。这需要根据患者的心理和行为特征，评估、建模并个性化制定干预措施，确保将心理健康作为肥胖症管理的基本组成部分。

然而，在实践中，我们对伴有心理健康共病的肥胖患者的理解，受到零散、稀疏且存在偏倚的真实世界临床数据的限制。许多队列对严重精神疾病表型的代表性不足，或缺乏一致的心理评估；缺失模态、编码不一致以及数据共享限制进一步制约了对代谢和心理健康维度的联合建模。合成数据提供了一种可行的方式，用于生成能够保留人口统计学特征、临床、行为和社会心理变量之间统计学上真实关系的人工队列，从而在保护数据隐私的同时，填补数据空缺、进行特征插值，并以计算机模拟方式研究罕见或极端表型 [5]。此外，在医学领域，合成队列有助于开展假设检验、模型原型开发和跨站点验证，包括针对真实数据集很少涵盖的低频但临床上至关重要的情况进行压力测试 [6]。

为解决这些局限，我们提出了一种多智能体系统（MAS），旨在模拟肥胖研究中的真实患者。该系统整合了多源证据，包括从医疗理赔数据中提取的诊断轨迹、干预措施和合并症病史，以及来自临床文献的见解和从人口调查中获得的行为、生活方式及社会心理分布，例如美国疾病控制与预防中心（CDC）的全国健康与营养调查（NHANES）和行为风险因素监测系统（BRFSS）。此外，我们还纳入人格维度、六因素人格模型（涵盖诚实—谦逊、情绪性、外向性、宜人性、尽责性和经验开放性（HEXACO）） [9],五因素强化敏感性理论（RST） [10], 以及气质与性格量表（TCI） [11]，为每位模拟患者赋予丰富的人格量表，以调节其行为反应、情绪反应性和治疗依从性。这种方法保留了代谢、

*aaghaee@klick.com
†sasgarian@klick.com
‡cjeon@klick.com

心理和人格相关特征之间的联合依赖关系，从而能够系统地探索人格特质与肥胖—心理健康诊断轨迹及治疗结果之间的关联。通过这种方式，本研究整合了临床、心理和行为数据流，模拟出更真实的肥胖队列。通过将多智能体建模与生成式合成相结合，本研究为肥胖护理领域可扩展、保护隐私且兼顾心理学视角的研究奠定了基础。

2 相关工作

2.1 模拟真实患者的早期方法

早期医疗患者模拟方法采用基于规则的模拟引擎。一个突出的例子是 Synthea，它利用模块化的疾病和工作流脚本模拟患者的完整生命周期，并生成结构化的电子健康记录（EHR） [12]。后续改进引入了统计校准和多类型事件建模，以提升人口统计学特征和流行病学的真实性 [13, 14]。这些基于规则的系统仍然很有价值，因为它们生成的记录可解释、符合架构规范，并与医疗标准一致。然而，其设计在很大程度上是确定性的，因此难以捕捉高维时间变化和细微的共病模式。

2.2 用于患者模拟的深度生成模型

为了直接从真实 EHR 中学习复杂的时间依赖关系和多模态依赖关系，研究人员采用了深度生成模型。SynTEG 和 EVA 等早期框架引入了变分自编码器和对抗自编码器，用于生成纵向诊断序列，并通过“合成数据训练、真实数据测试”（TSTR）评估来量化保真度和效用 [15, 16]。EHR-M-GAN 将这一思路扩展至连续型与类别型混合的重症监护病房（ICU）时间序列，证明合成数据可以提升预测性能 [17]。较新的基于 Transformer 和扩散模型的方法通过建模不规则的时间间隔和异质数据类型，进一步提升了真实性 [18, 19]。总体而言，这些深度模型在保真度和效用方面远胜于基于规则的系统，但在表征临床推理过程以及不同角色之间的多智能体交互方面仍有局限。

2.3 大语言模型、多智能体系统与临床叙述

近年来，大语言模型（LLM）取得的进展推动合成电子健康记录研究转向生成叙事内容丰富、具备上下文感知能力的记录，以及智能体式临床模拟。基于提示的方法将结构化电子健康记录序列化为文本，以进行条件生成 [20], ；而将结构化诊疗记录与单次就诊级或纵向笔记相结合的方法，则能提升上下文连贯性和下游文本分类的实用性 [21]。多角色生成式智能体由临床医生、患者和编码员大语言模型协作生成临床文档，标志着生成式电子健康记录向智能体式生成迈出了概念性的一步 [22]。然而，许多此类系统仍以文本为中心，缺乏完整的结构化代码输出、标准规范执行，以及定量的隐私与效用审计。既有精神科模拟框架（例如面向心理健康的对话式智能体）展现了对话的真实性，却未能保证纵向数据的保真度 [23, 24]。

3 方法 s

3.1 数据

本研究的最终目标是模拟各个维度都真实可信的患者，涵盖人口统计学特征、行为特征和临床多样性。因此，我们在多智能体系统框架中整合了多个互补数据集。这些数据源共同提供群体层面分布、纵向疾病模式和病例层面的临床细节。整合采用分层策略：（i）使用全国性调查数据建立人口统计学特征和行为基线；（ii）使用医疗理赔数据集捕捉纵向疾病轨迹；（iii）使用流行病学数据进行概率校准；（iv）使用患者病例报告实现临床和心理现实感。这种多模态设计使多智能体系统（MAS）能够模拟在人口统计学上具有代表性、在临床上连贯且以实证数据为基础的患者。

3.1.1 NHANES

我们使用了美国疾病控制与预防中心的全国健康与营养调查（NHANES）中的12个调查周期（1999–2023） [7]。这些调查结合访谈、体格检查和实验室检测，评估具有全国代表性的美国人口的健康与营养状况。我们提取的合并数据集涵盖21个健康领域，包括人口统计学因素、人体测量指标、慢性疾病和代谢性疾病、生活方式行为、心理健康指标、临床管理，以及生殖健康和文化适应等专门主题。各周期的数据经过标准化，并通过唯一受访者标识符进行关联，从而构建适用于后续综合分析的协调化参与者级记录。

3.1.2 医疗理赔数据

我们从 PurpleLab 医疗理赔数据库（https://purplelab.com）中提取了 70,000 名去标识化患者的数据，这些患者在过去 10 年的回溯期内至少有一条肥胖相关理赔记录。每位患者的纵向记录均包含人口统计学特征、诊断、治疗操作和药物信息，从而能够重建详细的疾病与治疗轨迹，适用于时间建模。

3.1.3 流行病学数据

肥胖及选定心理合并症的人群层面概率源自多个权威来源，包括美国疾病控制与预防中心的行为风险因素监测系统数据 [8], 、世界肥胖联合会 [25], 以及全国共病调查复制研究 [26]。这些分布为人口统计属性（年龄、性别、种族、教育程度、职业、收入）以及关注的目标共病（暴食症、抑郁症、焦虑症和社交恐惧症）设定了抽样先验，从而确保模拟队列能够反映美国人群数据中观察到的真实肥胖患病率和共病模式。

3.1.4 患者病例报告

我们从生物医学与生命科学文献数据库 PubMed 收集了患者病例报告。这些报告主要发表于经过同行评审的临床和医学期刊，包含详细的临床叙述，涵盖诊断、治疗及心理社会信息。这些内容有助于针对特定疾病提取相关的患者层面证据。基于这一语料库，我们整理了与分配给每位模拟患者的疾病相匹配的病例研究。检索到的叙述为症状模式、行为特征和治疗历程提供了实证依据，随后用于提升模拟患者档案的真实性和医学连贯性。

3.2 使用多智能体系统生成合成患者

我们提出的方法遵循一个多阶段框架，旨在模拟真实且连贯的患者记录。该流程始于数据预处理：将包括美国国家健康与营养调查、医疗理赔数据、流行病学数据集和患者病例报告在内的多个来源进行标准化、协调并整合为统一的患者层面记录。预处理完成后，通过一个由五个不同步骤构成的协同多智能体系统开展患者模拟，每个步骤均由一个专用智能体负责。摘要智能体将实证数据浓缩为结构化输入；生成器智能体创建初始患者档案；增强智能体为其补充纵向信息和循证医学信息；评估智能体审核其合理性和一致性；优化器智能体则解决发现的问题，生成最终的经验证档案。图1展示了整体工作流程。

3.2.1 摘要智能体

摘要智能体通过为每位模拟患者构建基础档案，启动多智能体系统框架。此过程将流行病学概率与现实世界的临床数据相结合，以确保起始设定真实可信。该智能体首先以概率方式生成人口统计学特征，并从指定的共病组中分配一种特定状况（即单纯肥胖，或肥胖合并暴食症、焦虑症、抑郁症或社交恐惧症）仅肥胖，或肥胖同时伴有暴食症、焦虑症、抑郁症或社交恐惧症）。

为使这一画像以真实数据为依据，该智能体从两个不同来源获取数据：一个包含 NHANES 数据中 10 个代表性个体样本的资料库，以及一个包含来自医疗理赔数据的 20 份纵向患者记录的数据库。随后，该智能体使用匹配算法，找出与生成的人口统计学特征和合并症画像最为匹配的单个 NHANES 样本。接着，它从医疗理赔数据中选出临床历程最符合生成画像与所选 NHANES 样本综合特征的前三项。

最后，该智能体对这些信息进行总结，生成一份结构化蓝图，为后续的生成器智能体提供指引。该蓝图包含患者的人口统计学特征、特定的体质指数范围（I级、II级或III级）、指定的合并症、匹配到的 NHANES 样本摘要，以及所选三项疾病历程的简要总结。此输出构成了模拟完整且逼真患者记录的基础框架。

3.2.2 生成器智能体

摘要智能体通过汇总参考样本，确定每位患者的档案及其潜在健康轨迹后，生成器智能体便将这一蓝图转化为完整、逼真的模拟。它将结构化摘要转化为内容完整、临床逻辑连贯的模拟患者档案，读起来如同真实的病历。在年龄、体质指数和合并症模式等人口统计学特征与临床参数的引导下，该智能体补全患者的背景、状况和经历等所有细节。

每份档案都通过受控的推理过程生成，以确保准确性和一致性。该智能体构建若干关键部分，共同呈现患者完整的健康历程：人口统计学特征和病史、当前状况、症状、习惯及实验室指标。它还描述治疗计划、心理特征和行为倾向，这些因素会影响患者的动机和依从性。为使每份病历都显得真实，该智能体还会纳入疾病时间线，追溯患者数年来的历程，展现其生理和心理因素如何随时间演变。

3.2.3 增强智能体

生成器智能体模拟出基础患者档案后，增强智能体负责为模拟补充循证的临床和行为细节，以提升其医学真实性。该智能体的主要功能是修订患者记录，整合直接从现实世界医学文献中提取的细致研究发现。

此流程首先识别生成的患者档案中所有相关的疾病关键词。随后，使用这些关键词查询 PubMed 数据库，为每个关键词检索 10 份相关的真实患者病例报告。接着，摘要智能体会处理这些文献，并

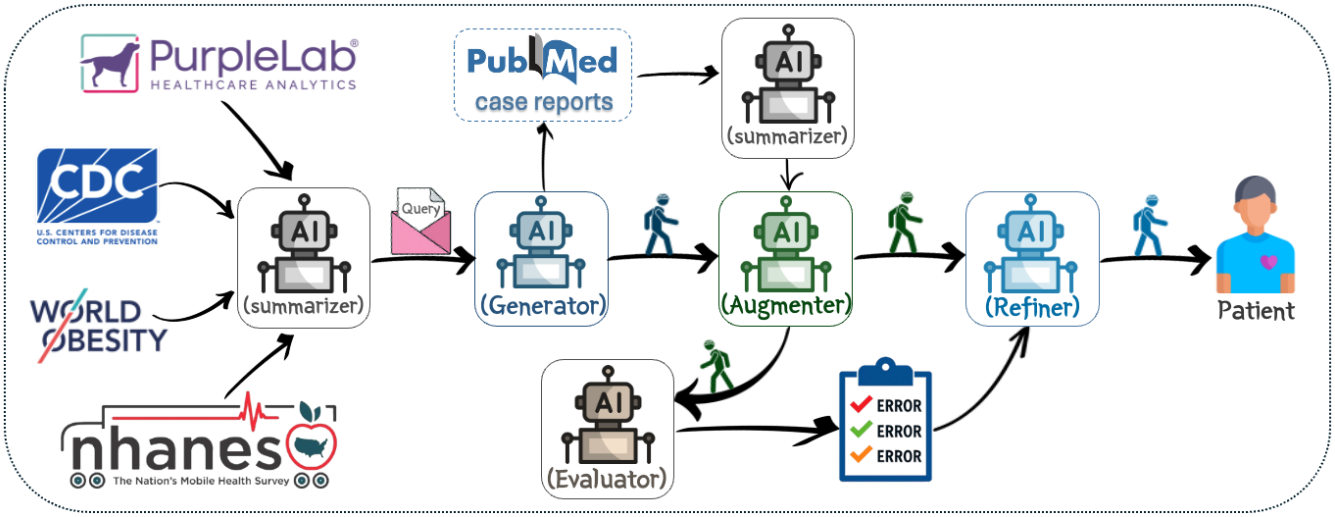

图1： SynthAgent 肥胖患者模拟多智能体框架概览。该系统整合了多个实证来源的数据，包括 NHANES、医疗理赔数据、流行病学数据集和患者病例报告。经过数据预处理后，摘要智能体将实证证据浓缩为结构化输入，供生成器智能体创建初始患者档案。增强智能体利用基于文献的临床证据丰富这些档案，而评估智能体则检查其中是否存在逻辑、时间或行为不一致。随后，优化器智能体解决所有问题，生成最终经过验证的合成患者。

首先筛选检索结果，排除所有与模拟患者的人口统计学特征（尤其是年龄和性别）不符的病例。随后，它会分析剩余的病例报告，综合总结文献记录的疾病通常如何影响患者的经历、症状、行为模式和心理状态。

随后，这些综合临床证据会传递给增强智能体，同时生成器智能体也会向其提供基础患者档案。增强智能体会细致地将这些见解融入档案，并有针对性地完善症状、心理量表和角色扮演档案等部分。例如，如果综合证据表明焦虑症与睡眠障碍之间存在强相关性，这一细节便会融入合成患者的行为叙述和症状表现中。通过以真实患者病例的证据为每份记录提供依据，这一过程弥合了统计真实性与真实人性细节之间的差距，最终生成一名经过增强的合成患者：不仅符合临床预期，还呈现出真实个体所特有的细腻情感与行为复杂性。

3.2.4 评估智能体

模拟患者档案经过医学信息补充后，评估智能体便会充当自动化临床审查员，确保每条记录都合乎逻辑、真实可信且内部一致。它以虚拟质量审计员的视角审查完整档案，检查患者生活与健康的多个维度中是否存在错误或矛盾。

智能体会审查人口统计学详情，确认职业、收入水平和保险类型与个人的年龄和背景相符。它还会检查医疗信息，核实疾病、症状和治疗是否恰当，并确保不存在不可能的组合或时间顺序错误，例如治疗出现在诊断之前。智能体还会评估心理连贯性，确保人格评分和行为特征与患者报告的习惯及情绪模式相符。最后，智能体会评估生活方式真实性，确认饮食、活动和睡眠行为与临床表现一致。

完成审查后，评估智能体会编制一份报告，总结发现的所有问题，并将其严重程度评定为严重、中等或轻微。这份报告将指导框架的下一阶段，帮助优化器智能体生成最终的模拟患者档案。

3.2.5 优化器智能体

优化器智能体是多智能体系统框架的最后一个阶段，负责确保每位模拟患者的记录准确、连贯且可供使用。它如同临床数据专家，审阅评估智能体提供的报告，并对模拟患者档案进行精确调整，以修正其中的问题。它不会重新生成整份记录，而是谨慎地只修改必要内容，以维护患者故事的完整性，并确保人口统计、临床和心理维度之间保持一致。例如，修正时间线不一致、调整不匹配的诊断，或完善行为细节。如果没有发现严重问题，该智能体会进行最后一次质量检查，以发现细微的不一致或遗漏的细节。最终得到的是一份经过润色和验证的模拟患者档案，忠实呈现真实的健康状况和行为动态。

3.3 模拟患者档案评估

我们评估了多智能体系统框架，以衡量生成的合成患者的真实性、连贯性和临床质量。为公平评估不同核心推理引擎对最终模拟质量的影响，我们建立了四种不同的多智能体系统配置。此设计的依据是保持输入不变，同时改变负责复杂模拟任务的模型。为此，所有实验均使用由 Gemini 2.0 Flash 驱动的摘要智能体。该智能体生成了 30 份标准化的基础档案，作为四种流水线配置完全相同的输入，从而确保比较过程受到控制。

区分四种多智能体系统配置的主要变量，是用作整个核心智能体套件推理引擎的大语言模型：生成器、增强器、评估器和精炼器。我们测试了四种在这一角色中领先的大语言模型引擎：GPT-5 [28], Gemini 2.5 Pro [29], Claude 4.5 Sonnet [30], 和 DeepSeek-R1 [31]。

目标是在保持队列多样性的同时确保高质量。因此，评估结合了两种互补分析：（i）自动化质量评估，采用大语言模型即评判者框架作为评审者，根据连贯性、真实性和完整性进行评分；（ii）潜在多样性可视化，分析模拟患者之间基于嵌入的差异。

在自动化质量评估中，我们实施了大语言模型即评判者框架，利用 GPT-4o [32]作为独立评估器，对全部 120 份模拟患者档案进行一致且近似人类的判断。我们使用涵盖所有维度的详细评分标准，对每份档案进行系统评分，以评估其对医学准确性和叙事连贯性的体现：（1）人口统计学特征、（2）病史、（3）当前状况、（4）症状、（5）习惯、（6）实验室指标、（7）治疗、（8）心理量表、（9）角色扮演档案，以及（10）疾病时间线。

评分采用演绎式方案，从满分100分开始，针对三个核心标准中发现的具体错误扣分：信息充分性，用于判断细节是否全面且比例适当；逻辑一致性，用于核查档案不同部分之间的内在连贯性（例如，确保心理状况与行为模式相符）；以及医学合理性，用于确认临床细节是否切合实际（例如，治疗是否适用于所述状况和症状）。这一严格的流程为每位模拟患者生成了各维度得分和最终总体质量评分，从而支持细致且全面的评估。

为比较模型表现，我们汇总所有分数，并采用描述性统计（均值、中位数、标准差和极差）进行测量。我们使用学生t检验进行模型两两比较，以确定模型表现是否存在统计学差异。当 $p < 0.05$时，我们认为结果具有统计学显著性。

为评估.....的语义多样性与独特性模拟队列，我们对由多智能体系统框架的不同配置（即四种大语言模型引擎）生成的模拟患者队列进行了全面的基于嵌入的分析。采用Jina Embeddings v4 [33] 将患者记录转换为高维语义表示（768个维度）。人口统计信息被排除在队列多样性分析之外，以确保多样性度量反映的是临床与行为内容的差异，而非人口统计学特征的变化，从而能够公平比较总体患者记录多样性。接下来，我们通过测量每个患者嵌入到其各自MAS配置队列质心的平均欧几里得距离，计算了各MAS配置内部的MAS配置多样性评分。MAS配置相似度则使用余弦相似度和成对距离进行评估。为便于可视化，我们应用t-SNE [34] 将嵌入投影到二维空间，并将患者表示为分段圆，以表明其共病特征（肥胖合并抑郁症、焦虑症、社交恐惧症、暴食症，或单纯肥胖）。

这项全面评估提供了一种客观、可扩展的方法，用于衡量模拟患者数据的连贯性和临床保真度，同时兼顾队列多样性，从而能够在多智能体系统框架中对不同语言模型骨干进行定量比较。

4 结果

4.1 定性评估

我们评估了四种大语言模型（GPT-5、Claude 4.5 Sonnet、Gemini 2.5 Pro 和 DeepSeek-R1），将其作为多智能体系统框架中的患者模拟引擎。在 120 名模拟患者（每种引擎生成 30 名）中，我们发现 GPT-5 的平均质量评分最高（76.27 ± 2.84），Claude 4.5 Sonnet 紧随其后（76.17 ± 3.90）；后者生成了评分最高的单个合成患者（得分为 = 82; 表1）。这两种领先模型在患者模拟方面表现相近（$p = 0.91$）。相比之下，Gemini 2.5 Pro（71.2 ± 3.0）和 DeepSeek-R1（67.8 ± 3.3）的表现显著较低（表2）。

表1： 大语言模型引擎的总体性能评估。

LLM	均值 ±标准差	Min	Max	中位数
GPT-5	76.27 ± 2.84	72	81	76.0
Claude	76.17 ± 3.90	69	82	76.0
Gemini	71.17 ± 2.96	65	77	71.0
DeepSeek	67.83 ± 3.26	63	76	68.0

维度级评估进一步印证了我们先前的观察。尽管在这一细粒度层面上，GPT-5 和 Claude 4.5 Sonnet 总体表现优于 Gemini 2.5 Pro 和 DeepSeek-R1，但 DeepSeek-R1 在模拟患者当前状况方面表现尤为出色。GPT-5 在技术性和结构性类别中表现更佳，包括病史、治疗、角色扮演档案和疾病

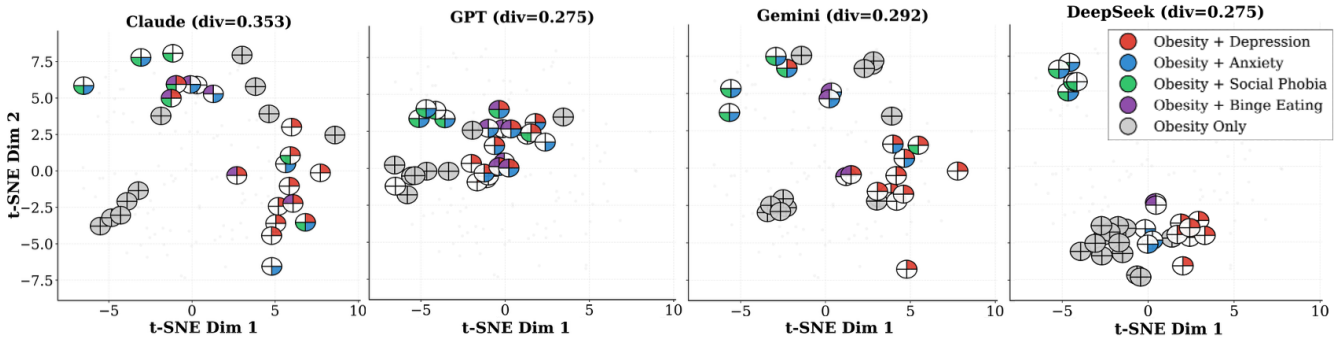

图 2： 四种大语言模型的临床核心嵌入 t-SNE 可视化。分段圆圈表示共病模式（红色=抑郁症、蓝色=焦虑症、绿色=社交恐惧症、紫色=暴食、灰色=仅肥胖）。

表2： 大语言模型引擎的两两比较。

比较	p 值	解释
Claude 与 GPT-5	0.910	等效
Claude 与 Gemini	<0.001	Claude ≫ Gemini
Claude 与 DeepSeek	<0.001	Claude ≫ DeepSeek
GPT-5 对比 Gemini	<0.001	GPT-5 ≫ Gemini
GPT-5 对比 DeepSeek	<0.001	GPT-5 ≫ DeepSeek
Gemini 对比 DeepSeek	0.0001	Gemini ≫ DeepSeek

时间线。相反，Claude 4.5 Sonnet 在症状、习惯、实验室检查和心理量表等维度上表现出色。这些发现表明，GPT-5 的优势在于结构连贯性和时间连贯性，而 Claude 4.5 Sonnet 则能提供更丰富的心理现实感。详细的维度级比较见表3。

表3： 大语言模型引擎各维度的质量评分。

维度	Claude	GPT-5	Gemini	DeepSeek
病史	99.6	**100.0**	95.8	95.8
当前状况	84.9	84.4	84.4	**86.2**
症状	**79.1**	76.2	73.6	70.7
习惯	**61.7**	53.8	52.5	43.3
Labs	**61.1**	52.3	52.3	59.0
治疗	70.9	**80.0**	73.3	65.1
心理量表	**71.3**	67.7	53.3	49.3
角色扮演	57.0	**60.7**	50.3	48.7
时间线	81.0	**82.3**	77.2	69.5

4.2 队列多样性分析

基于嵌入的分析揭示了四种不同多智能体系统配置所生成的模拟患者队列在语义多样性方面存在显著差异。Claude 4.5 Sonnet 核心引擎生成的患者档案最为多样，获得了最高的多样性评分（div = 0.353）。相比之下，DeepSeek-R1 和 GPT-5 生成的队列在语义上更加同质化，表现为较低的多样性评分（两者均为 div = 0.275）。

这些定量结果通过患者嵌入的二维 t-SNE 投影得到了直观印证。如图 2 所示，每个圆圈代表一名通过模拟生成的肥胖患者，彩色扇区表示特定的合并症，例如抑郁症或焦虑症。这些圆圈在二维空间中的排列反映了它们在语义上的相似性。我们观察到，DeepSeek-R1 和 Gemini 2.5 Pro 生成的患者，尤其是单一状况档案（仅肥胖），呈现出密集聚类，表明各自队列内部具有高度的语义相似性。相比之下，GPT-5，尤其是 Claude 4.5 Sonnet 生成的患者嵌入，在语义空间中的分布更加广泛。Claude 4.5 Sonnet 队列的离散程度尤为明显，即使是具有完全相同合并症档案的患者之间，也表现出清晰的区分度和丰富的变异性。这证实了其在本次模拟中生成多样且各具特色的临床叙述的卓越能力。

4.3 案例研究

我们提供一位模拟患者的详细资料作为案例研究，以展示我们的方法（图 3）。在我们的模拟中，患者档案显示，这是一名居住在美国北卡罗来纳州、已婚的 51 岁男性，患有 I 级肥胖（体质指数 33），并伴有高血压、高脂血症、阻塞性睡眠呼吸暂停、胃食管反流病和骨关节炎等合并症，以及失眠等有记录的心理和行为特征。全面的实验室检查结果（糖化血红蛋白 5.3%、空腹血糖 95 mg/dL、总胆固醇 185 mg/dL、低密度脂蛋白 105 mg/dL、高密度脂蛋白 44 mg/dL、甘油三酯 160 mg/dL）与超重或 I 级肥胖患者及合并心理／行为障碍者已知的代谢特征相符。例如，对精神障碍病情稳定患者的研究发现，在这类人群中，体质指数与甘油三酯和低密度脂蛋白胆固醇呈正相关 [35]。此外，记录在案的心理特征，如注意力缺陷／多动障碍特征、睡眠障碍和行为习惯，反映了肥胖与心理健康负担之间已得到充分证实的关联（即肥胖人群出现抑郁症、焦虑症或进食障碍的可能性更高）[36]。这份档案内容丰富，涵盖人口统计学特征、病史、症状、习惯、心理量表以及

疾病时间线表明，所提出的多智能体系统框架成功生成了一位结构连贯、临床可信的模拟患者。总分为 81，进一步证实该档案达到了多维患者模拟的高质量标准。

5 讨论

5.1 我们的贡献

多智能体患者模拟框架的出现，标志着医疗健康研究在克服数据稀缺和碎片化方面迈出了变革性的一步。传统临床数据集往往缺乏理解肥胖合并精神障碍等复杂共病所需的心理、行为和情境丰富性。通过整合多种数据来源（即医疗理赔数据、NHANES 调查和流行病学数据），SynthAgent 展示了协调一致的多智能体架构如何共同模拟真实患者的多因素特征。每个智能体都具备专门能力，从而构建出自适应模拟环境，既反映现实世界中的差异性，又能保护隐私。这种方法不仅支持可复现且符合伦理规范的安全实验，还能在临床数据集中鲜少出现的情景下开展假设检验，例如由情绪或认知应激源引发的治疗中断。

除了应用于肥胖领域，SynthAgent 还彰显了多智能体框架在推进以患者为中心的医疗建模方面的潜力。通过捕捉患者随时间变化的决策过程、动机转变和心理社会适应情况，这些系统为了解患者的亲身经历提供了视角，弥合了临床结局与行为动态之间的鸿沟。此类模拟队列能够提供对患者多样性的可解释、可检验表征，从而为数字疗法、精准行为干预措施和 AI驱动的决策支持工具提供依据。随着医疗人工智能转向更加以人为中心的模型，SynthAgent 这类多智能体患者模拟框架将在推动跨学科研究、训练新一代临床人工智能系统，以及重新构想如何将患者理解为动态适应系统而非静态数据点方面发挥关键作用。

5.2 患者模拟中大语言模型引擎的比较表现

结果表明，尽管 GPT-5 和 Claude 4.5 Sonnet 生成的合成患者总体质量在统计学上相当，但其底层模型展现出 distinct 且互补的优势。GPT-5 擅长生成结构和时间上连贯的患者叙事，其在病史、治疗和疾病时间线等维度上的优异表现便证明了这一点。这表明，在纵向一致性和历史准确性至关重要的模拟场景中，GPT-5 更为适用。相比之下，Claude 4.5 Sonnet 的突出优势在于生成具有更高心理真实性的语义多样化患者队列，其在症状、习惯和心理量表等维度上的表现优于其他模型。

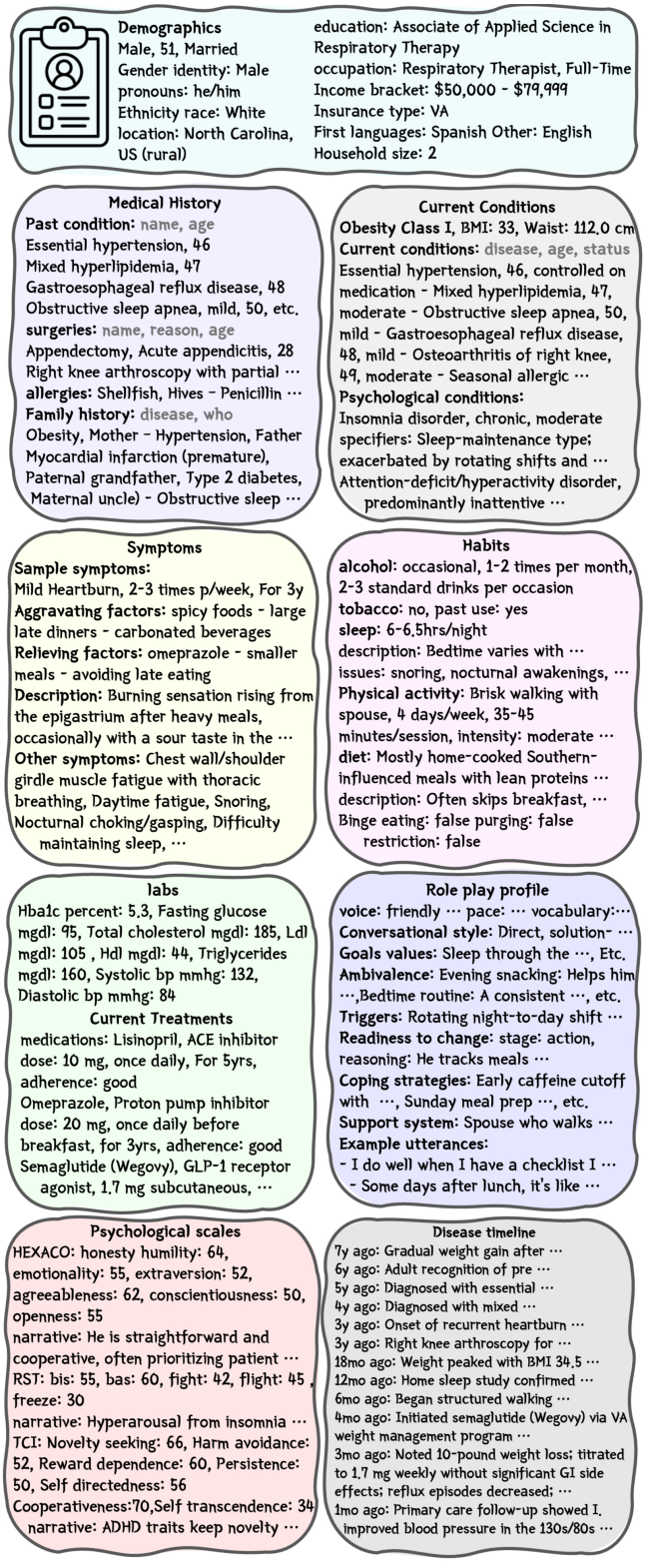

图 3： 使用 GPT-5大语言模型引擎生成的、由所提出的多智能体系统（MAS）创建的模拟患者示例。为适应可用空间，部分维度已截断，省略号（. . .）表示信息仍有后续内容。该患者的总分为 81。

Claude 4.5 Sonnet 生成的患者队列在语义上更加多样，并具有更高的心理真实性，其在症状、习惯和心理量表等维度上的表现优于其他模型。更为重要的是，Claude 4.5 Sonnet 的患者多样性显著高于 GPT-5；这种更丰富的多样性可以弥补质量评分上任何被认为轻微的差异，使其有望成为所提出的患者模拟多智能体系统的更优引擎。Gemini 2.5 Pro 和 DeepSeek-R1 的表现显著较低，表明它们不太适合这一任务，尽管 DeepSeek-R1 在描述患者当前状况方面展现出特定优势。总的来说，尽管 GPT-5 在结构保真度方面更具优势，但 Claude 4.5 Sonnet 在生成多样且心理层次丰富的患者队列方面已证实的实力，使其成为一个极具竞争力的选择。这些发现还表明，未来可以构建混合系统，利用不同模型的独特优势，生成更加稳健且逼真的模拟患者数据。

5.3 局限性

尽管 SynthAgent 展现了多智能体系统在模拟兼具医学与心理特征的患者方面的潜力，但仍存在若干局限。首先，该框架的性能取决于医疗理赔数据和人口调查等输入数据的质量与代表性，而这些数据可能带有固有偏差，或对罕见的精神疾病表型代表不足。其次，尽管多智能体系统框架支持智能体间的模块化协调，但智能体间推理和冲突解决仍部分依赖规则，限制了涌现行为的复杂性。第三，心理和行为建模依赖人格特质、调查衍生分布等代理变量，无法充分捕捉现实世界的多样性或社会文化细微差异。最后，目前的评估侧重于内部质量和语义多样性，而非外部临床验证。未来的工作应纳入临床医生参与评估、强化驱动的智能体适应以及跨领域验证，以确保模拟患者更忠实地反映真实患者的动态特征，并支持转化医疗研究。

6 结论

在本研究中，我们提出了SynthAgent，这是一个多智能体模拟框架，能够生成在医学、心理和行为层面均得到丰富刻画的患者档案。其能力以肥胖伴共病精神障碍作为代表性病症进行了展示。与单模型生成方法不同，SynthAgent采用协同智能体进行数据融合、行为模拟和质量评估，整合医疗理赔数据、临床证据和心理社会分布，以生成真实的患者轨迹。我们的研究结果凸显了多智能体合成患者生成在填补数据缺口、实现伦理安全实验，以及捕捉治疗依从性、情绪调节和生活方式适应的复杂模式方面的潜力。通过将医学真实性与心理深度相结合，SynthAgent为推进以患者为中心的建模奠定了可扩展的基础，并为理解不同病症下多样化的患者群体、决策过程以及健康管理行为提供了广泛的适用性。

参考文献

[1] Luppino, F. S.; de Wit, L. M.; Bouvy, P. F.; Stijnen, T.; Cuijpers, P.; Penninx, B. W.; and Zitman, F. G. 2010年。《超重、肥胖与抑郁症：纵向研究的系统综述与荟萃分析》。《普通精神病学档案》，67(3)：220–229。

[2] Duina, M.; Santi, G.; 和 Carraro, A. 2025年. 抑郁症患者身体活动的障碍与促进因素：社会生态模型框架下的系统综述. *Frontiersin Sports and Active Living*, 7: 1569335.[3] Trompeter, N.; Dârvariu, ‚S.; Brieva-Toloza, A. V.; Opitz, M.-C.; Rabelo-da Ponte, F. D.; Sharpe, H.; Desrivieres, S.; Schmidt, U.; 和 Micali, N. 2025年. 青少年焦虑症状与进食障碍症状之间的前瞻性关系：双向关系的系统综述与荟萃分析. 欧洲儿童与青少年精神病学, 34(6): 1691–1718.[4] Gerardo, G.; Peterson, N.; Goodpaster, K.; 和 Heinberg, L. 2025年. 抑郁症与肥胖. 当代肥胖报告, 14(1): 5.[5] Foraker, R.; Morrow, J. D.; Johnson, J. A.; Wilcox, A. B.; Forster, A. J.; 和 Payne, P. R. 2025年. 理解合成数据：面向真实世界证据的人工数据集. *BMJEvidence-Based Medicine*.[6] Gonzales, A.; Guruswamy, G.; 和 Smith, S. R. 2023. 医疗保健中的合成数据：叙述性综述. *PLOSDigital Health*, 2(1): e0000082.[7] NHANES. 2023. 国家健康与营养检查调查（NHANES）数据，1999–2023. https: // wwwn.cdc.gov/nchs/nhanes/. 美国马里兰州海厄茨维尔：美国卫生与公众服务部，美国疾病控制与预防中心.[8]BRFSS. 2024. 行为风险因素监测系统数据. 美国卫生与公众服务部，美国疾病控制与预防中心. 访问日期: 2025-10-20. 可获取于: https://www.cdc.gov/brfss/index.htm.[9]Wang, S.; Li, R.; Chen, X.; Yuan, Y.; Wong, D. F.; 和 Yang, M. 2025年. 探究人格特质对大语言模型偏见与毒性的影响. *arXiv* 预印本 *arXiv:2502.12566*.

[10] Krupi´c, D.; Corr, P. J.; Ruˇcevi´c, S.; Križani´c, V.; 和 Graˇcanin, A. 2016. 五份强化敏感性理论（RST）人格问卷：比较、效度与推广. 人格与个体差异, 97: 19–24.

[11] Tomita, T.; Aoyama, H.; Kitamura, T.; Sekiguchi, C.; Murai, T.; and Matsuda, T. 2000. 人格心理生物学七因素模型的因素结构：模型修订。人格与个体差异，29(4): 709–727.

[12] Walonoski, J.; Kramer, M.; Nichols, J.; Quina, A.; Moe- sel, C.; Hall, D.; Duffett, C.; Dube, K.; Gallagher, T.; and McLachlan, S. 2018. Synthea：一种生成合成患者及合成电子医疗记录的方法、流程和软件机制。《美国医学信息学会杂志》，25(3): 230–238.
[13] Zhang, Z.; Yan, C.; Mesa, D. A.; Sun, J.; and Malin, B. A. 2020. 通过改进训练、建模和评估，确保电子病历模拟的可靠性。《美国医学信息学会杂志》，27(1): 99–108.[14] Yan, C.; Zhang, Z.; Nyemba, S.; and Malin, B. A. 2021. 生成具有多种数据类型和约束的电子健康记录。载于*AMIA*年度研讨会论文集，第2020卷，1335.[15] Zhang, Z.; Yan, C.; Lasko, T. A.; Sun, J.; and Malin, B. A. 2021. SynTEG：用于模拟时间结构化电子健康数据的框架。《美国医学信息学会杂志》，28(3): 596–604.[16] Biswal, S.; Ghosh, S.; Duke, J.; Malin, B.; Stewart, W.; Xiao, C.; and Sun, J. 2021. EVA：使用条件变分自编码器生成纵向电子健康记录。载于医疗保健机器学习会议，260–282。PMLR.[17] Li, J.; Cairns, B.; Li, J.; and Zhu, T. 2023. 为人工智能应用生成合成混合类型纵向电子健康记录。《*NPJ*数字医学》，6(1): 98.[18] Yuan, H.; Zhou, S.; and Yu, S. 2023. EHRDiff：探索使用扩散模型合成逼真的电子健康记录。arXiv。*arXiv*预印本 *arXiv:2303.05656*.[19] Karami, H.; Atienza, D.; and Ionescu, A. 2024. Synehrgy：使用仅解码器Transformer合成混合类型结构化电子健康记录。*arXiv*预印本 *arXiv:2411.13428*.
[20] Wang, Z.; and Sun, J. 2024. PromptEHR：基于提示学习的条件式电子医疗记录生成。” arXiv，2022年10月11日。访问日期：*5*月23日。

[21] Sun, H.; Lin, H.; and Yan, R. 2024. 通过多次就诊健康状态推断协同合成患者记录。载于*AAAI*人工智能会议论文集，第38卷，19044–19052页。[22]Velzen, M. v.; van der Willigen, R. F.; De Beer, V. J.; de Graaf-Waar, H. I.; De Graaf-Waar, E. R.; van Leeuwen, E.; van der Willigen, M. F.; Van Der Willi-gen, M. J.; Renardus, G.; El Maaroufi, R.; 等。2025。通过生成式智能体实现隐私、语言和信息保真的临床文档合成。《人工智能前沿》，8：1644084。[23] Wang, R.; Milani, S.; Chiu, J.; Zhi, J.; Eack, S.; Labrum, T.; Murphy, S.; Jones, N.; Hardy, K.; Shen, H.; 等。2024。PATIENT-Ψ：利用大语言模型模拟患者，以培训心理健康专业人员（arXiv：2405.19660）。arXiv。[24] Lee, J.; Lim, K.; Jung, Y.-C.; and Kim, B.-H. 2025。Psyche：用于评估精神科评估对话智能体的多面向患者模拟框架。*arXiv*预印本 *arXiv:2501.01594*。[25] 全球肥胖观察站。2025。美国：流行病学数据概览。https://data.wor ldobesity.org/country/united-states- 227/。全球肥胖观察站。伦敦：世界肥胖联合会。[26] Kessler, R. C.; and Merikangas, K. R. 2004。全国合并症调查复制研究（NCS-R）：背景与目标。《国际精神病学研究方法杂志》，13（2）：60–68。[27] Ashwell, M.; and Gibson, S. 2016。腰高比作为“早期健康风险”指标：比使用基于体质指数和腰围的“矩阵”更简单且预测性更强。《*BMJ*开放》，6（3）：e010159。[28] OpenAI。2025。GPT-5。https://openai.com/gpt-5/。大语言模型。[29] Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; 等。2025。Gemini 2.5：凭借高级推理、多模态、长上下文和新一代智能体能力推动前沿发展。*arXiv*预印本 *arXiv:2507.06261*。[30]Anthropic。2025。Claude Sonnet 4.5。
h
ttps://
www.anthropic.com/news/claude-sonnet-4-5。大语言模型。[31] DeepSeek-AI。2025。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。*arXiv*预印本 *arXiv:2501.12948*。大语言模型。[32] OpenAI。2024。GPT-4o系统卡。*arXiv*预印本 *arXiv:2410.21276*。

[33] Günther,M.; Sturua, S.;Akram,M.K.;Mohr,I.; Ungure-anu, A.; Wang, B.; Eslami, S.; Martens, S.; Werk, M.; Wang, N.; 等。2025。jina-embeddings-v4：用于多模态多语言检索的通用嵌入。*arXiv*预印本 *arXiv:2506.18902*。[34] Maaten, L. v. d.; and Hinton, G. 2008。使用t-SNE可视化数据。机器学习研究期刊，9（11月）：2579–2605。[35] Li, X.; Shi, X.; Tan, Y.; Yu, Y.; Tang, C.; Xu, G.; Zhang, X.; Liao, H.; Mai, X.; Chen, W.; 等。2022。稳定期常见精神障碍患者的肥胖代谢指标。*BMC*精神病学，22（1）：91。[36] Segal, Y.; and Gunturu, S. 2024。与肥胖相关的心理问题。载于*StatPearls [Internet]*。StatPearls Publishing。