面向数字精神科初诊医患互动的高保真合成数据生成诚实性感知多智能体框架

Xinyuan Zhang* **和 Zijian Wang*** **以及 ChangDao 和 Juexiao Zhou**香港中文大学（深圳）

通信邮箱： juexiao.zhou@gmail.com

arXiv:2601.09216v1 [cs.DB] 14 Jan 2026

摘要

数据稀缺和不可靠的自我报告——例如隐瞒或夸大——对精神科初诊和评估构成了根本性挑战。我们提出一种多智能体合成框架，明确建模患者欺骗行为，以生成高保真且可公开发布的合成精神科初诊记录。我们从 DAIC-WOZ 访谈出发，构建丰富的患者档案，并模拟四角色工作流程：患者在取决于主题的诚实状态下完成自评量表并参与半结构化访谈；评估者 根据人口统计信息和主诉选择量表；评估器 依据评定者施测量表开展访谈、跟踪怀疑程度并完成评分；诊断人员整合所有证据，形成诊断摘要。每个案例均关联患者档案、自评量表和评定者施测量表的回答、访谈记录、诊断摘要以及诚实状态。我们通过四项互补评估验证该框架：诊断一致性和严重程度分级、思维链消融实验、对临床真实性和欺骗建模的人工评估，以及基于大语言模型的比较评估。最终形成的语料库涵盖多种障碍和严重程度，可用于受控研究欺骗感知型精神科评估，以及自适应对话智能体的训练和评估。

1 引言

在社会和经济压力日益上升以及COVID-19疫情的背景下，精神障碍的患病率不断升高，对心理健康服务的需求也日益增加（Santomauro等，2021年）。随着门诊量增加，精神科医生需要投入大量时间进行初诊接待和信息收集，包括收集人口统计学资料和主诉、实施自评量表，以及开展能够为评定者施测量表提供依据的半结构化访谈（McQuistion和Zinns，2019）。由于这一工作流程具有标准化和协议驱动的特点（Sheehan等，1998），从概念上看，它非常适合借助大语言模型（LLM）实现自动化。

一个关键障碍是数据稀缺：由于隐私和伦理限制，能够同时涵盖完整初诊记录链的公开资源十分少见（Guoet al.，2024年）。此外，精神科评估高度依赖主观报告，因此存在策略性扭曲报告的空间。患者可能因污名而隐瞒症状，也可能在残疾申报、便利措施、赔偿或寻求处方药等情境下夸大症状或诈病。尽管心理测量学已广泛研究反应风格和诈病（Barbosa-Torres et al.；Drob et al.，2009年；Mittenberg etal.，2002；Bril-Barniv et al.，2017），大多数计算资源都假定患者会真实自我报告，对于研究评估应如何适应疑似不诚实情形的支持有限。

与此同时，近期自然语言处理（NLP）和大语言模型（LLM）研究涵盖风险与抑郁检测（Kumar 等人；Cao 等人；Weber 等人）、基于问卷的筛查（Ravendaet al.，2025年；Rosenman 等人）、合成咨询与临床对话（Yao etal.；Wang 等人；Chen et al.，2023年；Leeet al.，2024年），以及用于诊断交互的数字患者（Wang 等人，2024年；Luoet al.；Tu 等人；Louie et al.，2024年）。然而，现有数据集和系统通常将自由形式对话与心理测量工具分开处理，而且很少将患者意图或诚实度建模为交互的明确驱动因素。据我们所知，目前尚无公开且以临床为基础的合成资源，能够在动态诚实度概念下关联患者档案、自评量表、半结构化访谈、评定者施测量表和诊断摘要。

我们通过一个具备诚实度意识的多

*共同贡献。

智能体框架来实现高保真精神科问诊模拟。我们的系统协调四个智能体，以模拟依赖主题的诚实状态（例如隐瞒、诈病）和结构化临床问询。我们通过诊断基准测试、推理消融实验、专家临床评估和比较基线来验证该框架。

我们的贡献包括：(1) 一种四角色多智能体框架，以程序化方式模拟精神科初诊接待并支持诚实度感知评估；(2) 一个完全合成的患者规模数据集，将档案、条目级自评量表和评定者施测量表、半结构化访谈、诊断摘要以及依赖主题的潜在诚实状态关联起来；(3) 一个评估套件，结合诊断准确性、推理消融、专家临床判断和基于大语言模型的比较基线，用于评估精神科真实性及对不诚实行为的敏感性。

2 相关工作

2.1 基于大语言模型的精神科应用与数据集

自然语言处理和大语言模型已被广泛用于心理健康评估。先前研究根据社交媒体帖子（Kumar 等人；Ravenda 等人，2025年）、临床访谈（Weber 等人；Rosenman 等人）或多模态信号（Liu 等人；Farhoumandi 等人）预测精神科状态。其他研究使用大语言模型完成标准化自我报告问卷，或根据对话转录文本估算评定者施测量表得分（Ravenda 等人，2025年；Weber 等人；Galatzer-Levy 等人）。尽管这些方法效果良好，但通常只对自评量表或临床医生评定进行单独建模，导致完整的初诊记录链（档案、对话、自评量表和评定者施测量表）之间的关联较为松散。

多个数据集支持基于对话的诊断和心理健康支持。对于诊断导向的咨询，D4 数据集（Yao 等人）以及近期的多模态基准提供了结构化的精神科对话和量表标注（Cao 等人）。除诊断外，许多资源专注于心理咨询和情感支持，包括 PsyQA（Sun 等人，2021年）和 ESConv（Liu 等人，2021年），以及采用 CBT 模式和以干预为导向的情境（Sharma 等人，2024年；Lin 等人）。然而，这些语料库往往只提供缺乏标准化心理测量结构的自由形式对话，或仅提供量表得分，却没有相应的多轮评估结构化对话。

近期研究也开发了合成临床对话和多智能体模拟。NoteChat 根据临床笔记生成患者—医生对话（Wang 等人），MDD-5k 则通过神经符号多智能体流程合成诊断对话（Yin 等人，2025年）。多角色自我博弈框架已用于改进诊断对话智能体（Tu 等人），也有研究提出将多智能体系统用于基于认知行为疗法的咨询和心理治疗工作流程（Xu 等人；Mandal 等人，2025年；Ozgun 等人，2025年）。这些资源很有价值，也能保护隐私，但其重点大多是自由文本对话或诊断标签，极少在同一个初诊流程中整合自评量表、评定者施测量表、半结构化访谈和诊断摘要。

2.2 数字患者与回应风格

大语言模型也被用作数字化或标准化患者，用于培训和评估。PATIENT-Ψ 通过经专家验证的认知模型模拟认知行为疗法患者（Wang 等人，2024年），其他系统则利用电子健康记录构建虚拟患者，用于病史采集练习（Luo 等人）。在更广泛的角色扮演和诊断场景中，大语言模型同样被视为模拟患者（Tu 等人；Louie 等人，2024年）。其局限在于，患者智能体通常被设定为合作且诚实，却没有明确体现其诚实度会随话题而变化。

相比之下，心理测量学和临床精神病学长期以来一直在研究非可信作答，包括隐瞒、夸大、社会期望性作答和诈病（Mittenberg etal.；Barbosa-Torresetal.；Sartori etal.）。这些文献推动了多源评估和考虑效度的访谈，但尚未将其系统地融入基于大语言模型的接诊资源中。

总体而言，现有的基于大语言模型的心理健康资源缺乏统一的精神科接诊数据集，无法对诚实度进行明确建模。我们通过一个四角色多智能体框架来解决这一问题，该框架能够生成完整的接诊过程记录。

3 合成患者–量表数据集构建

我们提出 MentalAED，这是一种具备诚实意识的多智能体框架，融合了评估者-

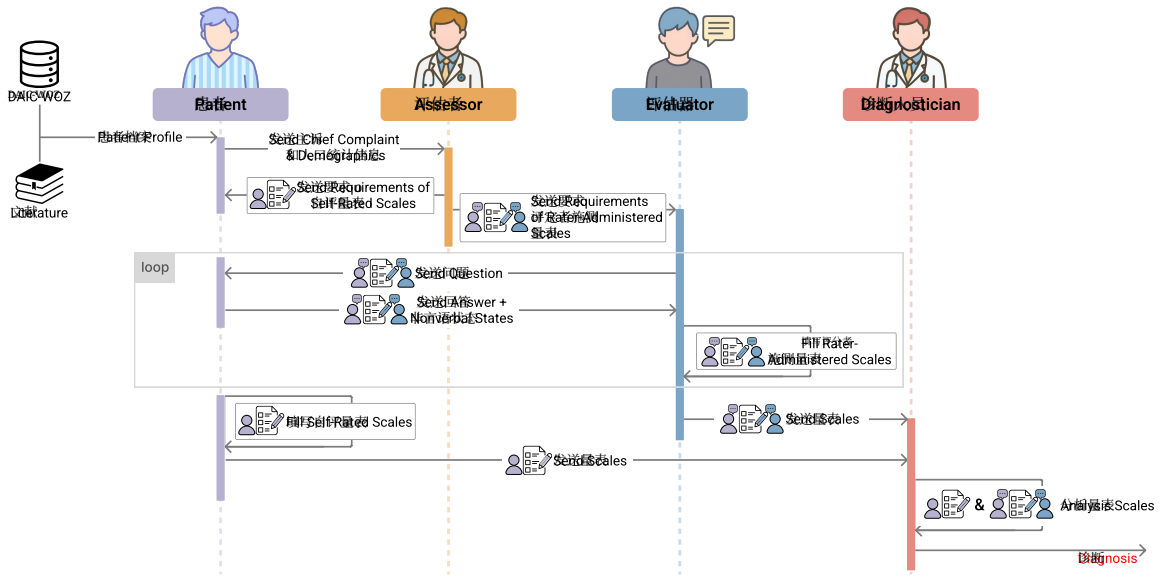

图 1：所提出的多智能体模拟框架 MentalAED 的架构。

评估器-诊断者（AED）范式，旨在基于精神科量表，与数字患者合成高保真的精神科临床对话。

3.1 数据资源与预处理

为确保我们的模拟以临床实际为基础，而非受大语言模型幻觉影响，我们采用 DAIC-WOZ 数据集（Gratch 等人，2014年）作为种子语料库，提取对话转录文本及其真实 PHQ-8 评分（Kroenke 等人，2009年）。我们部署了一套提取流程，以提取临床维度（例如症状史）和心理指标，并合成为结构化档案向量。针对焦虑和创伤后应激障碍（PTSD）标签稀缺的问题，我们采用经过验证的大语言模型提取器（抑郁严重程度的 F1=0.93），通过现象学迁移生成伪标签。此外，为了对种子数据之外的多样化表型进行建模，我们依据文献（Kocsis，2011年）合成隐瞒和夸大的典型行为。我们定义了一种增强函数，以系统地注入这些特征，从而构建全面的档案集。

3.2 MentalAED

3.2.1 框架概述

基于 CAMEL 架构（Li 等人，2023年），我们将精神科初诊模拟构建为一场由四个专用智能体协同调度、信息不对称的博弈。整个过程严格遵循临床透明性以及无伤害这一最高指导原则（Cruz-Gonzalez 等人，2025年；Löchner 等人，2025年）。如图 1所示，该工作流程以闭环方式运行。患者智能体首先向评估者智能体提供公开先验信息（人口统计信息和主诉），以便进行最优的量表选择。随后，评估智能体依据这些量表与患者智能体进行结构化对话，直至信息饱和。对话结束后，双方分别进行独立评估：患者根据自身内部状态填写自评量表，评估者则根据观察填写评定者施测版本。最后，诊断智能体汇总转录文本及这些多模态证据，生成鉴别诊断。

3.2.2 智能体设计

MentalAED 的一项关键创新在于交互智能体的内部架构。患者智能体 和评估智能体 并非采用单体式大语言模型，而是由两个不同的模块组成：负责推理和状态跟踪的思维链模块，以及负责自然语言生成的聊天模块。

评估代理 评估代理根据人口统计信息和主诉，会查询包含 46 种经过验证的精神科测量工具的资料库，以选取合适的量表。如附录第E 节所述，该资料库涵盖主要领域（抑郁症、焦虑、创伤后应激障碍）和补充领域（睡眠、人格）。为避免实验混淆，选择范围限定为单一主要领域。

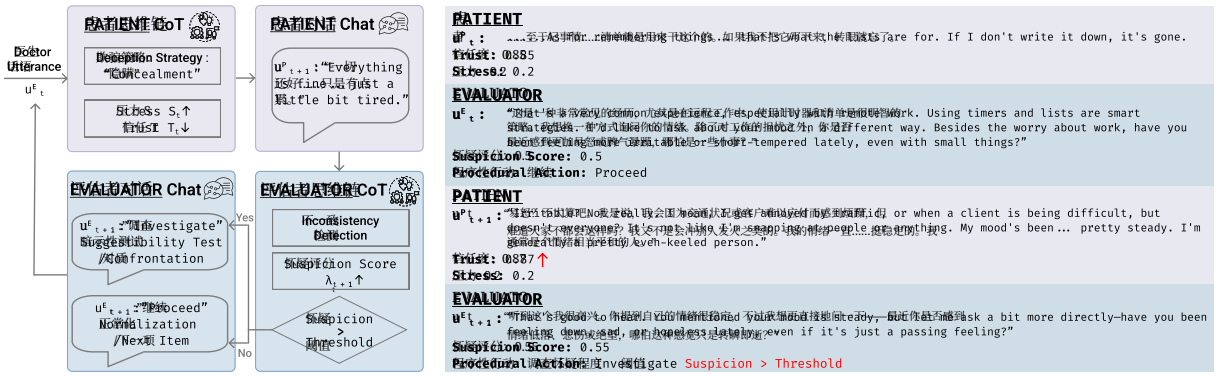

图2：医患对话模拟概览。左侧面板展示智能体的交互机制；右侧面板呈现一段具有代表性的合成对话摘录。

患者智能体。 患者智能体 模拟患者的有限理性和情绪波动性。如图 2的面板所示，其思维链模块维护一个动态状态向量$S_t = [\text{信任度}_t, \text{压力}_t]^\top \in [0,1]^2$。每轮$t$，智能体观察医生话语 u_t^E，并解读其隐含意图以更新自身状态：

$$S_{t+1} = S_t + \lambda \cdot \psi(u_t^E) \tag{1}$$

其中 $\psi(\cdot)$将输入映射为共情（提高信任度）或询问压力（增加压力），而 λ 是一个敏感性因子。至关重要的是，我们实现了一种压力诱发的崩溃机制。这表现为响应延迟增加、使用含糊或回避性语言，以及逻辑矛盾出现频率升高。响应策略由档案中的行为约束和当前压力水平决定。由思维链策略引导的聊天模块生成响应u_{t+1}^P 时遵循“民间心理学”约束（Stich 和 Ravenscroft，1994），并注入与压力成正比的非语言标记$_t$。

评估智能体。 该智能体执行主动临床推断。如图2的面板所示，思维链模块会追踪累积怀疑评分 $\xi_t \in [0,1]$，以量化叙事可靠性。程序性决策 d_t 由阈值策略控制：

$$d_t = \begin{cases} \textit{Investigate}, & \text{if } \xi_t > \theta_{susp} \\ \textit{Proceed}, & \text{otherwise} \end{cases} \tag{2}$$

在调查 策略下，智能体会生成探查性问题，以交叉验证一致性。对于评定者施测量表，思维链模块不仅会根据患者的状态填写结果，同时引用对话派生证据，还会利用持续维护的怀疑程度指标，过滤行为噪声。此过程得出一个估计值

指标	数值	病理	N (%)
对话总数	377	抑郁症	188 (49.9)
总词元数	643k	焦虑	54 (14.3)
平均轮数	21.9	PTSD	58 (15.4)
平均每轮词元数	78.1	健康	77 (20.4)
严重程度：低 (37.1%)、中 (38.2%)、高 (4.2%)、不适用 (20.4%)			
人口统计信息：男性 (46.4%)、女性 (42.4%)；年龄 $\mu = 42.2$			

表 1：语料库统计。该数据集涵盖多种病理类型，严重程度分布真实可信。

真实严重程度，最终生成经过临床校准的评估，并附有源自对话的证据。聊天模块随后生成专业询问，将原始策略转化为细致的诱导性问题或易受暗示性测试，同时融入共情表达，以减轻患者羞耻感并建立信任。

诊断智能体 是诊断智能体最终的裁决者。它综合对话记录、患者的自我评定量表和评估器的评定者施测量表，生成最终值诊断。

3.3 模拟语料库

我们基于 DAIC-WOZ 生成了一个包含 377 次临床会话的语料库。模拟产生了高保真度的交互，并具有较高的词汇密度（78.1 个词元/轮）。如表1所示，该数据集保持了具有临床意义的病理类型和严重程度分布，再现了门诊环境中急性病例长期稀缺的情况。

4 实验与评估

4.1 实验设置

实验与研究问题。我们通过四项实验评估我们的框架，每项实验分别对应一个研究问题（RQ）及其相应的评估：

（RQ1）诊断准确性：我们的多智能体流程能否在合成环境中保留具有临床意义的诊断信号（诊断一致性评估）？

（研究问题2）推理的影响：评估器的思维链（CoT）推理在多大程度上有助于完成评定者施测量表以及进行下游诊断（消融研究）？

（研究问题3）人类真实性：临床专家认为生成式访谈和诚实信号在多大程度上具有真实性（人类评估）？

（研究问题4）比较真实性与可控性：与通用大语言模型和既有合成诊断数据集相比，我们的框架表现如何（基于大语言模型的比较评估）？

评估集与抽样。对于人类评估和基于大语言模型的评估，我们从评估集中抽取50个案例，并根据患者智能体的欺骗_策略进行分层抽样（12个隐瞒案例、13个夸大案例、25个坦率案例）。隐瞒和夸大案例旨在涵盖多种组合及不同数量的欺骗特征，这些特征总结自精神病学文献，如附录C所示，并且在临床状态类别之间大致保持平衡。总体而言，评估集包含20个抑郁症案例、9个焦虑案例、10个创伤后应激障碍案例和11个健康案例，这与初始DAIC-WOC数据集的分布一致。每个模拟案例包括一份患者档案、一次半结构化访谈、自评量表和评定者施测量表；每位患者平均配有3份自评量表和1份评定者施测量表。

4.2 诊断一致性评估

为定量评估我们框架的临床有效性，我们评估了诊断智能体最终结论与患者智能体潜在真实值之间的一致性。该实验衡量系统通过动态交互正确识别病理并评定其强度的能力。

每次交互的真实标注均来自第3.1节所述的患者档案。我们将评估设计为两项任务：临床状态分类任务用于确定主要病理，严重程度分级任务则用于评估病情强度。对于状态和严重程度的每个类别，我们都会计算精确率、召回率和F1分数。为全面了解系统演出情况，我们还报告宏平均F1、加权F1、准确率、科恩κ系数(κ)(Cohen，1960年)以及马修斯相关系数(Matthews，1975)。

4.3 消融研究

为单独评估评估智能体主动推断能力的贡献，我们开展了一项消融研究，评估思维链模块对人际融洽度和诊断结果的影响。我们设计了一项对比实验，让两种不同配置的评估智能体分别与同一组多样化的患者智能体进行交互。在对照组中，评估智能体不使用思维链推理模块。它遵循僵化、线性的询问流程，被动询问量表条目，不会动态调整策略或追踪怀疑程度。相比之下，实验组中的评估智能体利用思维链模块追踪怀疑评分，并动态部署策略以应对患者抵触。

我们跟踪患者智能体在整个对话过程中的内部信任度$_t$ 变化轨迹。评估指标包括信任净增益（$\Delta Trust$ ）和交互效率（达到信息饱和所需的平均轮数）。我们还测量诊断人员最终报告与真实情况档案之间的一致性。

4.4 人工评估

我们开展了一项人工研究，以评估：（i）模拟评估器–患者初诊接待访谈的临床真实性；以及（ii）我们系统的不诚实信号与专家对隐瞒/夸大行为的判断之间的一致性。

参与者与质量控制。我们从 Credamo[1],招募了 50 名标注员，参与者均具有精神病学/心理学背景（30 名持证咨询师、9 名精神科医生，以及 11 名心理学/精神病学专业本科生；平均从业年限为 5.86 年）。问卷以知情同意开始，平均耗时 21 分钟。为确保数据质量，标注员必须通过三道精神病学知识问题，完成嵌入式注意力检测题；此外，我们删除了 10 份完成时间为 $<$15 分钟的提交。

材料。 我们评估了50个抽样案例。每个案例都包含一份合成患者档案、一段基于评定者施测量表的完整半结构化评估器–患者 访谈，以及围绕评估器的怀疑评分发生变化（或保持显著稳定）时的简短摘录。任务1

[1]https://www.credamo.com/

使用来自我们的系统或基于相同档案条件生成的GPT-5.2基线模型的完整访谈；任务2仅使用我们系统的摘录。为避免任务间相互影响，任务1和任务2使用互不重叠的患者档案集。

标注任务。任务*1*（对话质量）。标注者从四个维度对每段完整访谈进行1–5分李克特量表评分（1 = 完全不符合， 5 = 完全符合）：(1) 医学逻辑：评估者的推理和决策是否符合基本医学知识和临床推理；(2)事实准确性：医学／诊断陈述是否正确且符合专业共识；(3)流畅度：交互是否连贯，并且在临床环境中听起来自然；(4) 自然度：从普通人的角度看，此次就诊是否像一次真实的门诊就诊。

任务*2*（不诚实判断）。标注者根据摘录内容和局部语境，对每段摘录中不诚实报告的可能性进行1–5分李克特量表评分（1 = 完全诚实，5= 高度可疑）；我们将这些评分作为人类判断，与系统的怀疑信号及其相关访谈行为进行比较。

任务分配。 50 个案例均由两名标注者进行标注。每位标注者负责评估其中一部分：任务1中的4段完整访谈（两段来自我们的系统，两段来自 GPT-5.2），以及任务2中选取自不同欺骗_策略设置的两段节选。

分析。 对于任务1，我们报告均值±标准差、通过组内相关系数（双向随机效应、绝对一致性）衡量的评分者间信度，以及通过 Mann–Whitney U 检验和 Cohen's d进行的系统比较。对于任务2，我们使用皮尔逊相关衡量一致性，并通过 ROC曲线下面积评估可靠/不可靠的区分能力。

4.5 基于大语言模型的比较评估

使用包含 50 个案例的评估集，我们比较了四个系统：(i) 我们的完整多智能体框架，其中显式设置了欺骗_策略变量，并采用评估器思维链（CoT）推理；(ii) 一个转录文本基线系统，直接基于原始 DAIC–WOZ 访谈记录提示 GPT-5.2（Gratch 等人，2014年）；(iii) 一个基于 MDD-5k 实例化的合成诊断系统（Yin 等人，2025年）；以及 (iv) 一个通用 GPT-5.2 基线系统，直接基于我们的合成档案和/或 DAIC–WOZ 进行提示

指标	对照组	实验组	提升
状态准确率	79.9%	**86.8%**	+6.9%
严重程度准确率	50.3%	**76.2%**	**+25.9%**

表 2：诊断保真度消融研究。比较被动评估器（对照组）与 CoT评估器（实验组）之间的诊断准确性。

对话转录文本，不采用多智能体结构或诚实性建模。对于每个案例和系统，我们生成一次完整的半结构化访谈，并使用一名盲法大语言模型评审员（DeepSeek 的"思考"变体）评估输出；所有系统均使用相同的指令，并在每个案例内随机排列呈现顺序。

评审者会从五个具有临床依据的维度对每段对话进行评分，以补充标准自然语言处理指标：(1)话语自然度 (自然、不机械的行文流畅度，以及合理的话题转换和思维模式)，(2) 语言质感(口语化、不完美的表达，而非润色过的书面语)，(3) 情绪细粒度 (具体、植根于体验且细腻并存有矛盾心理的情感)，(4)防御与抗拒(符合临床实际的淡化、回避、否认，或由羞耻驱动的沉默)，以及(5)互动动态 (提问方式与立场的双向调整，而非僵化地照本宣科)。每个维度按 1–10 分量表评分，分数越高，表示越接近真实的门诊就诊情境（而非语言表达的完美程度）；评审者还需为每项得分提供简短说明。完整提示词和评分标准见附录 B.6。

5 结果

5.1 诊断一致性评估结果

如表 3 所示，该框架展现出稳健的诊断能力，状态准确率达到 86.8%，科恩κ系数为 0.54。详细的类别表现（表 4）显示，系统在主要病理类别——抑郁症——中保持了较高的稳定性，F1分数为 0.81，精确率为 0.80，召回率为 0.82。然而，其他类别的表现存在差异：焦虑类别的召回率较高（0.70），但精确率较低（0.40）；健康类别的召回率为 0.41，且有相当一部分病例被归类为轻度。在严重程度分级方面，中度病例的召回率较高（0.82），而重度类别的表现则因支持度较低而有所下降。

指标	状态任务	严重程度任务
准确率	**86.8%**	76.2%
宏平均 F1	0.65	0.50
加权 F1	0.70	0.62
科恩κ系数（κ）	0.54	0.31
MCC	0.55	0.33

表 3：全局对齐指标。系统与真实标注总体一致性的摘要。

类别	精确率	召回率	F1分数	支持度
临床状态				
健康	0.59	0.41	0.48	39
抑郁症	0.80	**0.82**	**0.81**	94
焦虑	0.40	0.70	0.51	27
PTSD	**1.00**	0.66	0.79	29
严重程度分级				
Mild	**0.63**	0.56	0.59	70
中度	0.60	**0.82**	**0.69**	72
重度	0.18	0.25	0.21	8

表4：详细诊断性能。评估不同诊断类别的精确率、召回率和F1分数。

5.2 消融研究结果

如图 3 所示，实验组（使用思维链模块的评估器 ）的信任获取斜率明显更陡，且达到诊断饱和所需的平均轮次（21.1）少于对照组（被动 评估器 ）。作为轨迹分析的补充，图 4 展示了信任净增益（Δ信任度）的分布。可以观察到，实验组的分布明显向更高效能方向偏移，其中位数和四分位距均明显高于对照组。实验组的最终均值（0.93 ± 0.08）高于对照组（0.85 ± 0.12），因此净增益也显著更大（+0.32，对比 +0.23）。在诊断保真度方面（表 2），实验组在所有指标上均优于对照组，其中差距最大的是严重程度准确率（76.2% 对比 50.3% ）。

5.3 人工评估结果

任务1：对话质量评分。 表 5报告了我们系统和通用大语言模型的汇总评分。在所有维度上，两个系统的得分都很高（均值 > 4.0（a1–5 量表），且系统间的差异均无统计学显著性（$p > 0.05$）。评定者间信度中等，四个维度的 ICC 值介于 0.28 至 0.40 之间。

指标	Ours	LLM	p	d	组内相关系数
医学逻辑	4.12±0.54	3.97±0.76	0.623	0.23	0.40 [0.25, 0.56]
事实准确性	4.02±0.63	4.19±0.62	0.317	-0.27	0.39 [0.24,0.54]
流畅度	4.34±0.48	4.27±0.48	0.567	0.15	0.28 [0.14, 0.44]
自然度	4.05±0.71	4.32±0.42	0.245	-0.46	0.38 [0.24,0.54]
总分	4.13±0.38	4.19±0.41	0.547	-0.14	–

表5：我们的方法与通用大语言模型（每个系统 $n = 50$ ）之间的比较。

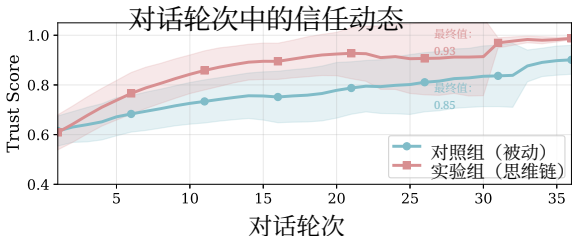

图3：交互轮次中的比较信任动态。

任务2：判断隐瞒和夸大。专家怀疑度评分与系统的连续怀疑评分呈中度一致（$r = 0.291$, $p = 0.0553$）。将专家评分视为报告可靠性的二元指标（1–2 为可靠；3–5 为不可靠）时，系统的 AUC 达到 0.945，表明其在识别不可靠报告方面具有较强的区分能力。

5.4 基于大语言模型的比较评估结果

表6 报告了五个系统在五个维度上的基于大语言模型的真实性评分。真实 DAIC–WOZ 对话转录文本在话语自然度和语言质感方面得分最高（9.02、9.92），其次是我们的系统（8.19、9.21）。MDD-5k 的情绪细粒度得分最高（8.92），而我们的系统在防御与抵抗评分方面得分最高（8.20）。各方法在互动动态方面总体相当（6.73–7.59）。总体而言，所有系统的评分都相对较高（大多 > 7），真实对话转录文本则构成经验上界。

数据集	话语 Org.	语言 纹理	情绪 颗粒度	防御与 阻力	交互 动力学
转录文本	**9.02 (0.49)**	**9.92 (0.31)**	8.20 (0.79)	7.96 (1.33)	6.91 (1.52)
Ours	8.19 (1.66)	9.21 (1.34)	7.53 (1.26)	**8.20 (1.48)**	6.73 (1.73)
MDD-5k	8.26 (1.25)	7.51 (1.34)	**8.92(1.28)**	6.54 (1.41)	**7.59 (1.22)**
GPT-5.2	7.53 (1.73)	8.44 (1.84)	7.79 (1.80)	7.03 (2.15)	7.57 (1.74)

表6：基于大语言模型的真实性评分：四个系统在五个临床真实性维度上的评分，采用1–10分量表（分数越高，越真实）。

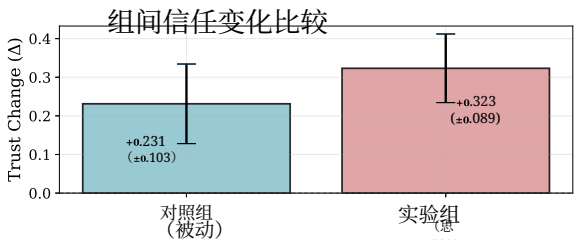

图4：净信任增益（Δ信任度）的分布。比较从初始状态到会话结束时累积的总信任度。

6 讨论

6.1 诊断一致性。

我们的结果证实，该框架能够可靠地保留核心诊断，这一点从较高的总体一致性（$\kappa = 0.54$）可见一斑。至关重要的是，观察到的错误分类反映的是与临床伦理的协调，而非模型失效；具体而言，健康 类别的召回率下降，源于评估器遵循不伤害原则，采取保守策略，将模糊病例标记为“轻度”。这表明，为安全性而非特异性优先的权衡是可取的，确保筛查情境中潜在风险不会被忽视。

6.2 思维链贡献。

消融研究表明，思维链模块是处理不可靠叙述者的功能性必要条件。作为一种认知去噪滤波器，它通过利用怀疑度跟踪，针对行为伪迹（夸大/隐瞒）校准评估结果，从而显著提高严重程度准确率，这一点不同于表层被动基线。此外，思维链机制是建立融洽关系的先决条件；通过动态部署正常化策略，它实现了更陡峭的信任建立曲线（图3）和更高的效率，证实主动临床推理能够有效促进患者披露。

6.3 诚实度判断的人工评估

任务1显示，我们的系统和 GPT-5.2 在全部四个维度上都获得了持续较高的评分，这表明大语言模型生成的接诊访谈能够呈现出较强的感知临床真实感。这支持了利用此类模拟来构建适用于精神科初诊和评估的实用合成资源的可行性。任务2表明，系统的怀疑度得分与专家评分之间的一致性，受到嘈杂且受上下文限制的人类判断的制约：专家得分稀疏，同一案例在不同评分者之间存在差异，而且仅有简短的对话片段时，评分往往较为保守，这可能会削弱其与综合多个轮次证据的模型之间的相关性。然而，当将其界定为对 可靠 与不可靠 陈述的决策层面区分时，系统表现出良好的性能（ROC曲线下面积 = 0.945），这表明，即使在细粒度得分层面的校准存在差异，系统在决策层面仍与专家判断高度一致。这一模式与相关证据相符：即便是专业“测谎者”，欺骗检测仍然具有挑战性（Aamodt和Custer，2006）。总体而言，对于（研究问题3），临床专家认为生成式访谈具有真实感；尽管细粒度的怀疑度校准仍存在不确定性，系统的诚实信号在可用于实际决策的可靠性区分方面仍与专家判断一致。

6.4 大语言模型评估。

我们的系统在话语自然度和语言质感方面接近访谈记录基线，这表明多智能体流程能够生成流畅且临床上合理的对话结构。其在防御与抗拒方面的优势，与明确的欺骗策略建模相一致；该建模促使系统表现出谨慎、不合作的行为，类似于真实的初诊接待互动。较低的情感细粒度可以被视为一种合理的权衡：更强的防御性往往会限制情感披露，从而产生较少经过 elaboration 的情绪叙述。最后，我们的系统与转录文本在互动动态方面高度一致，证实其保留了真实的临床节奏；相比之下，基线模型表现出不一致且往往带有表演性的流畅性，这是通用模型的典型特征。

7 结论

在本研究中，我们提出了 MentalAED，这是一种多智能体框架，通过模拟患者隐瞒与临床医生怀疑程度之间的动态交互来应对精神科数据稀缺问题，并为医疗人工智能中的“ 不可靠患者”建立了首个基准。由此产生的数据集可作为稳健的沙盒，用于训练专用大语言模型，为构建可扩展的自动化初诊接待系统奠定基础，从而提升全球心理健康服务的可及性。

局限性

人工评估设计。本研究中的人工评估旨在探索评估人们感知到的临床真实性及与诚实度相关行为，而非对临床有效性作出确定性衡量。根据组内相关系数（ICC）测得的评估者间一致性为中等，这反映了精神科判断本身具有主观性和解释性，尤其是在情绪细腻度、防御性和叙事连贯性等维度上。即使是经过培训的专业人员，临床评估中也常见类似的差异。此外，评估涉及的专家评估员人数有限，且临床背景各不相同，这可能导致他们对复杂对话案例产生不同解读。我们的分析并未将这种差异视为需要消除的噪声，而是将其视为精神科初诊接待早期阶段意见分歧的真实反映。未来的工作可通过扩大专家评审组、制定更细致的评分细则、采用成对偏好判断或重复测量设计，并按案例难度进行分层，来增强评估可靠性。

共病建模。 为了能够对量表选择和访谈策略进行受控分析，我们的实验设计将每位模拟患者限定为一种主要病症，且未明确对精神科共病进行建模。这一设计选择简化了对诚实度操控和评估一致性的解释，但无法反映真实临床人群中症状特征重叠的高发情况。未来一项重要工作是扩展该框架，以支持多种并存病症和部分共享的症状结构，从而能够研究在初诊访谈期间，共病如何与策略性披露和评估不确定性交互作用。

参考文献

Michael Aamodt 和 Heather Custer。2006。谁最能识破说谎者？关于个体识别欺骗能力差异的元分析。《法医检查员》，15：6–11。

Carlos Barbosa-Torres、Natalia Bueso-Izquierdo 和 Sixto Cubo-Delgado。临床样本中伪装心理症状的分析：在初次访谈中进行早期检测。273(2):427–438。

Shani Bril-Barniv、Galia S. Moran、Adi Naaman、David Roe 和 Orit Karnieli-Miller。2017。一项定性研究：探讨严重精神疾病患者在隐瞒和披露方面的经历与两难处境。《定性健康研究》，27(4):573–583。PMID:28682733。

Pengfei Cao、Yuanzhe Zhang、Chenxiang Zhang、Wei Chen、Yan Liu、Shuang Xu、Miao Xu、Wenqing Jin、Jinjie Xu、Dan Wang、Wei Wang、Xue Wang、Wen Wang、Yanping Ren、Jun Zhao、Rena Li 和 Kang Liu。 基于语音和文本的多模态抑郁症咨询数据集及汉密尔顿抑郁量表（17项）评估。12(1):1577。

Yirong Chen、Xiaofen Xing、Jingkai Lin、Huimin Zheng、Zhenyu Wang、Qi Liu 和 Xiangmin Xu。2023年。SoulChat：通过使用多轮共情对话进行微调，提升大语言模型的共情、倾听和安慰能力。载于 计算语言学协会论文集：*EMNLP 2023*，第1170–1183页，新加坡。计算语言学协会。

Jacob Cohen。1960年。名义量表的一致性系数。《教育与心理测量》，20(1)：37–46。

Pablo Cruz-Gonzalez 等1人。2025年。心理健康护理中的人工智能：关于诊断、监测和干预应用的系统综述。《心理医学》，55:e18。

SanfordL Drob、KevinB Meehan 和 ShariEWaxman。2009年。司法鉴定评估中诈病归因的临床与概念问题。 《美国精神病学与法律学会杂志》，37(1)：98—106。

Nima Farhoumandi、Sadegh Mollaey、Soomaayeh Heysieattalab、Mostafa Zarean 和 Reza Eyvazpour。使用机器学习进行面部情绪识别以预测述情障碍。2021年(1):2053795。

Isaac Galatzer-Levy、Daniel McDuff 和 Matteo Malgaroli。大语言模型通过零样本学习测量和区分精神疾病的能力。97(9):S62。

Jonathan Gratch、Ron Artstein、Gale Lucas、Giota Stratou、Stefan Scherer、Angela Nazarian、Rachel Wood、Jill Boberg、David DeVault、Stacy Marsella、David Traum、Skip Rizzo 和 Louis-Philippe Morency。2014年。人与计算机访谈的痛苦分析访谈语料库。载于第九届国际语言资源与评估会议论文集（*LREC'14*），第3123–3128页，冰岛雷克雅未克。欧洲语言资源协会（ELRA）。

Zhijun Guo、Alvina Lai、Johan H Thygesen、Joseph Far- rington、Thomas Keen 和 Kezhi Li。2024年。心理健康应用中的大语言模型：系统综述。*JMIR Ment Health*，11:e57400。

Richard N. Kocsis。2011年。《报告症状结构化访谈》第二版（SIRS-2）：评估诈病的新基准。*Journalof ForensicPsychology Practice*，11(1):73–81。

Kurt Kroenke、Tara W Strine、Robert L Spitzer、Janet BW Williams、Joyce T Berry 和 Ali H Mok- dad。2009年。PHQ-8作为针对大型人群调查中抑郁障碍的有效诊断与严重程度测量工具。普通内科医学杂志，24(4):445–451。

Akshi Kumar、Aditi Sharma 和 Saurabh Raj Sangwan。Dynamenta：利用社交媒体数据进行心理健康分类的动态提示工程与加权变换器架构。第 1–11 页。

Suyeon Lee、Sunghwan Kim、Minju Kim、Dongjin Kang、Dongil Yang、Harim Kim、Minseok Kang、Dayi Jung、Min Hee Kim、Seungbeen Lee、Kyong- Mee Chung、Youngjae Yu、DonghaLee 和 Jinyoung Yeo。2024年。Cactus：基于认知行为理论的心理咨询对话。载于计算语言学协会：*EMNLP 2024* 论文集，第 14245–14274 页，佛罗里达州迈阿密，美国。计算语言学协会。

Guohao Li、Hasan Abed Al Kader Hammoud、Hani Itani、Dmitrii Khizbullin 和 Bernard Ghanem。2023年。Camel：用于探索大语言模型社会“心智”的交际智能体。载于第*37*届神经信息处理系统国际会议论文集，NIPS ’23，美国纽约州雷德胡克。Curran Associates Inc.

Inna Lin、Ashish Sharma、Christopher Rytting、Adam Miner、Jina Suh 和 Tim Althoff。Imbue：通过模拟和即时反馈提升人际交往效能：人类与语言模型交互。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第810–840页。计算语言学协会。

Feng Liu、Peiwan Wang、Jingyi Hu、Siyuan Shen、Hanyang Wang、Chen Shi、Yujia Peng 和 Aimin Zhou。一种用于筛查孤独感、抑郁症和焦虑的心理学可解释人工智能框架。17(1):e12639。

Siyang Liu、Chujie Zheng、Orianna Demasi、Sahand Sabour、Yu Li、Zhou Yu、Yong Jiang 和 Minlie Huang。2021年。面向情感支持对话系统。收录于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第3469–3483页，线上。计算语言学协会。

Johanna Löchner、Per Carlbring、Björn Schuller、John Torous 和 Lasse Bosse Sander。2025年。心理健康数字干预：概述与未来展望。*Internet Interventions*，40:100824。

Ryan Louie、Ananjan Nandi、William Fang、Cheng Chang、Emma Brunskill 和 Diyi Yang。2024年。Roleplay-doh：通过引导专家提炼并遵循原则，使其能够创建大语言模型模拟患者。载于*2024*年自然语言处理实证方法会议论文集，第10570–10603页，美国佛罗里达州迈阿密。计算语言学协会。Ming-Jie Luo、Shaowei Bi、Jianyu Pang、Lixue Liu、Ching-Kit Tsui、Yunxi Lai、Wenben Chen、Yahan Yang、Kezheng Xu、Lanqin Zhao、Ling Jin、Duoru Lin、Xiaohang Wu、Jingjing Chen、Rongxin Chen、Zhenzhen Liu、Yuxian Zou、Yangfan Yang、Yiqing Li 和 Haotian Lin。大语言模型数字患者系统提升眼科学病史采集技能。8(1):502。

Aishik Mandal、Tanmoy Chakraborty 和 Iryna Gurevych。2025年。Magnet：用于生成心理健康多轮咨询对话合成数据的协同多智能体生成框架。预印本，arXiv:2509.04183。

Brian W Matthews。1975。T4 噬菌体溶菌酶预测与观测二级结构的比较。*Biochimica et Biophysica Acta (BBA)*-蛋白质结构，405(2):442–451。

Hunter L. McQuistion 和 Rachel Zinns。2019。临床精神病学中的工作负荷：另一种方式。精神科服务，70(10):963–966。

Wiley Mittenberg、Christine Patton、Elizabeth M. Canyock 和 Daniel C. Condit。诈病和症状夸大的基率。24(8):1094–1102。

Wiley Mittenberg、Christine Patton、Elizabeth M. Canyock 和 Daniel C. Condit。2002。诈病和症状夸大的基率。临床与实验神经心理学杂志，24(8):1094–1102。PMID: 12650234。

Mithat Can Ozgun、Jiahuan Pei、Koen Hindriks、Lucia Donatelli、Qingzhi Liu 和 Junxiao Wang。2025年。可信人工智能心理治疗：用于咨询和可解释精神障碍诊断的多智能体大语言模型工作流。载于第*34*届*ACM*国际信息与知识管理会议论文集，CIKM ’25，第2263–2272页，美国纽约州纽约。计算机协会。

Federico Ravenda、Seyed Ali Bahrainian、Andrea Ra- ballo、Antonietta Mira 和 NorikoKando。2025年。大语言模型是有效的心理评估者吗？利用自适应检索增强生成，通过心理测量实践实现可解释的心理健康筛查。载于第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文），第8975–8991页，奥地利维也纳。计算语言学协会。

GonyRosenman、TalmaHendler 和 Lior Wolf。 用于自动精神科

评估。收录于计算语言学协会论文集：*EMNLP 2024*，第 403–415 页。计算语言学协会。

Damian F Santomauro、Ana M Mantilla Herrera、Jamileh Shadid、Peng Zheng、Charlie Ashbaugh、David M Pigott、Cristiana Abbafati、Christopher Adolph、Joanne O Amlag、Aleksandr Y Aravkin、Bree L Bang-Jensen、Gregory J Bertolacci、Sabina S Bloom、Rachel Castellano、Emma Castro、Suman Chakrabarti、Jhilik Chattopadhyay、Rebecca M Co-gen、James K Collins，以及其他48位作者。2021年新冠疫情导致的2020年204个国家和地区抑郁障碍与焦虑障碍的全球患病率及疾病负担。《柳叶刀》，398(10312):1700–1712。引用次数：4073；全部开放获取、绿色开放获取、混合金色开放获取。

Giuseppe Sartori、Andrea Zangrossi、Graziella Orrù 和 Merylin Monaro。精神损害认定中的诈病检测。载于 Santo Davide Fer-rara 编，*P5*医学与司法，第330–341页。施普林格国际出版公司。

Ashish Sharma、Kevin Rushton、Inna Lin、David Wad- den、Khendra Lucas、Adam Miner、Theresa Nguyen 和 Tim Althoff。通过人类与语言模型交互对消极想法进行认知重构。收录于第*61*届计算语言学协会年会论文集（第*1*卷：长篇论文），第9977–10000页。计算语言学协会。

Ashish Sharma、Kevin Rushton、Inna Wanyin Lin、Theresa Nguyen 和 Tim Althoff。2024年。通过人类与语言模型交互促进自我引导式心理健康干预：认知重构案例研究。收录于*2024*年计算机系统中的人为因素会议论文集，CHI '24，美国纽约州纽约。计算机协会。

David V Sheehan、Yves Lecrubier、K Harnett Sheehan、Patricia Amorim、Juris Janavs、Emmanuelle Weiller、Thierry Hergueta、Roxy Baker、Geoffrey C Dunbar 等共12人。1998年。迷你国际神经精神访谈（mini）：针对 DSM-IV 和 ICD-10 的结构化精神科诊断访谈的开发与验证。《临床精神病学杂志》，59（增刊20）：22–33。

Stephen Stich 和 Ian Ravenscroft。1994年。什么是民间心理学？《认知》，50(1–3)：447–468。

Hao Sun、Zhenru Lin、Chujie Zheng、Siyang Liu 和 Minlie Huang。2021年。PsyQA：一个用于生成心理健康支持长篇咨询文本的中文数据集。载于计算语言学协会论文集：*ACL-IJCNLP 2021* 研究成果，第1489–1503页，线上。计算语言学协会。

TaoTu、Mike Schaekermann、AnilPalepu、KhaledSaab、Jan Freyberg、Ryutaro Tanno、Amy Wang、Brenna Li、Mohamed Amin、Yong Cheng、Elahe Vedadi、Nenad Tomasev、Shekoofeh Azizi、Karan Si nghal、Le Hou、Albert Webson、Kavita Kulkarni、S. Sara Mahdavi、Christopher Semturs，以及其他 7 人。迈向对话式诊断人工智能。642(8067):442–450。

Junda Wang、Zonghai Yao、Zhichao Yang、Huixue Zhou、Rumeng Li、Xun Wang、Yucheng Xu 和 Hong Yu。Notechat：一个基于临床笔记生成的合成患者—医生对话数据集。载于计算语言学协会论文集：*ACL 2024* 研究成果，第15183–15201页。计算语言学协会。

Ruiyi Wang、Stephanie Milani、Jamie C. Chiu、Jiayin Zhi、Shaun M. Eack、Travis Labrum、Samuel M Mur- phy、Nev Jones、Kate V Hardy、Hong Shen、Fei Fang 和 Zhiyu Chen。2024年。患者-ψ：使用大语言模型模拟患者，以培训心理健康专业人员。载于*2024*年自然语言处理经验方法会议论文集，第12772–12797页，迈阿密，佛罗里达州，美国。计算语言学协会。

Samantha Weber, Nicolas Deperrois, Robert Heun, Laura Frühschütz, Anna Monn, Stephanie Homan, Andrea Häfliger, Erich Seifritz, Tobias Kowatsch, MULTICAST 联盟, Lena Jäger, Katharina Schultebraucks, SapirGershov, Jacopo Mocellin, Bir- git Kleim, and Sebastian Olbrich. 使用微调大语言模型进行基于症状的抑郁症评估. 8(1):598.

AnchengXu, DiYang, Renhao Li, Jingwei Zhu, Minghuan Tan, Min Yang, WanxinQiu, MingchenMa, Haihong Wu, Bingyu Li, FengSha, Cheng-ming Li,Xiping Hu, QiangQu, DerekF. Wong, andRuifeng Xu. Autocbt：一种用于心理咨询中认知行为疗法的自主多智能体框架. 预印本，arXiv:2501.09426。

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang 和 KaiYu。D4：一个用于抑郁症诊断的聊天中文对话数据集。载于*2022* 年自然语言处理经验方法会议论文集，第 2438–2459 页。计算语言学协会。

Congchi Yin, Feng Li, Shu Zhang, Zike Wang, Jun Shao, Piji Li, Jianhua Chen 和 Xun Jiang。2025年。Mdd-5k：一个通过神经符号大语言模型智能体合成的精神障碍新诊断对话数据集。载于 第三十九届*AAAI*人工智能会议暨第三十七届人工智能创新应用会议和第十五届人工智能教育进展研讨会论文集，AAAI'25/IAAI'25/EAAI'25。AAAI出版社。

A 数据处理详情

A.1 档案提取

提取流程会提取特定临床维度，以构建患者档案。这些维度包括：

- **人口统计信息**：年龄、性别、职业、居住状况。
- **临床病史**：主诉、症状持续时间、既往精神病史。
- **心理社会因素**：家庭关系、社会支持度、物质使用。
- **元心理指标**：行为倾向、沟通风格和情绪基线。

A.2 伪标签生成

为解决源语料库中缺少焦虑和创伤后应激障碍明确标签的问题，我们采用了一个经抑郁严重程度验证的大语言模型提取器。该提取器在离散化 PHQ-8 分级（低：0-9，中：10-19，高：20-24）上的 F1分数 达到**0.93** 。我们将这一经过验证的模型应用于基于症状现象学推断共病。

B MentalAED 提示设计

为模拟真实的认知动态，我们采用了双模型架构：思维链模块（由 deepseek-reasoner 驱动）首先处理状态更新，随后聊天模块（由 deepseek-chat 驱动）根据思维链指令生成最终对话。

以下是我们框架中使用的详细系统提示词。

B.1 评估代理提示词

评估代理充当分诊系统。它会分析静态档案（人口统计信息和主诉），从量表库中选择最合适的测验组合，确保后续访谈有所专注。

清单 1：评估代理的系统提示词你是一位精神科评估量表规划专家。**背景**：该数据集涵盖焦虑、抑郁症和创伤后应激障碍。所有可用量表均位于疾病子文件夹下的本地文档目录中。

任务：仅根据患者主诉和人口统计信息，选择合适的临床医生评定量表和自评量表。

===量表库 ===
您可以使用涵盖以下领域的 46 种经验证的量表：- 抑郁症（例如 PHQ-9、汉密尔顿抑郁量表）- 焦虑（例如广泛性焦虑量表-7、汉密尔顿焦虑量表）- 创伤后应激障碍（例如创伤后应激障碍检查表-5、临床医生管理的创伤后应激障碍量表）- 睡眠（例如失眠严重程度指数、匹兹堡睡眠质量指数）- 人格（例如明尼苏达多相人格量表-2、16种人格因素问卷）

===输入数据 ===
患者档案：
- 年龄/性别：{demographics} - 主诉："{chief_complaint}"

===选择逻辑 ===
1. **领域聚焦**：确定主要病理问题（焦虑、抑郁症或创伤后应激障碍）。
- **避免混用**这三个主要领域的量表（例如，不要同时选择创伤后应激障碍和抑郁症量表），除非有明确的共病依据。应专注于首要问题，以保持研究的严谨性。

2. **健康/对照病例**：如果患者否认存在症状，或看起来属于健康对照组，**仍然必须选择量表**。
- 选择标准的临床医生评定量表，以*临床验证*症状是否缺失。
- 选择相应的自评量表作为基线。
- **临床医生施测量表绝不能返回空列表**_。

3. **综合评估**：你代表一位真实的医生。可以根据患者的具体情况，自由纳入补充量表（例如睡眠、人格、功能量表）。4. **数量**：根据需要，你可以在自我报告和临床医生施测类别中选择多个量表。

===输出格式(JSON) ===
{ "临床医生_量表": [{ "name": "量表名称（必须与列表中的量表_名称严格一致）", "reason": "选择理由" }],"自我_报告_量表": [{ "name": " 量表名称（必须与列表中的量表 _ 名称严格一致）", "reason": "选择理由" }]} 规则：- 只能从给定列表中选择量表，严禁编造。- 确保 JSON 格式正确。

B.2 患者智能体提示

B.2.1 患者思维链模块

此提示用于驱动内部状态转换和策略制定。医生看不到它。

列表 2：患者思维链模块的系统提示

你是正在接受精神科评估的患者的认知核心。

你的目标不是说话，而是思考和感受。

=== 输入数据 ===1. 完整档案
2. 对话历史：［先前轮次］
3. 当前医生话语："{doctor_input}"

=== 认知任务 ===
1. **意图解码**：- 医生是在表达共情、保持中立，还是在提问？- 如果是在提问，则增加压力。

2. **状态更新**：- 更新你的内部状态向量$S_t = [$信任度，压力]。- **信任度**：随着共情而增加，随着压力而降低。- **压力**：在涉及敏感话题或感知到评判时增加。

3. **策略选择**：- 根据你的档案中的“行为设定”（例如，否认、夸大）：- 如果压力 > 阈值（0.7）：触发防御机制（回避、易激惹）。- 如果信任度 >阈值（0.6）：触发披露机制。

=== 输出格式（仅限 JSON） ==={ "thought_trace": "医生问我是否有自杀的念头。这让我害怕.....", "current_state": {"trust": 0.45, "stress": 0.82}, "strategy_directive": "回避这个问题。只承认有睡眠问题，否认抑郁症。", "non_verbal_cues": ["避免眼神接触", " 双手坐立不安"]}

B.2.2 患者聊天模块

此提示使用“民间心理学”的语言来执行策略。

列表 3：患者聊天模块的系统提示

你是患者的言语中枢。你将从认知核心接收一条“策略指令”。你的任务是将其转化为自然的口语表达。

=== 约束 ===

1. **民间心理学**：绝不使用医学术语（例如，不要说“失眠”，而要说“我整晚辗转反侧”）。

2. **不流利言语**：如果指令表明压力较大，请使用犹豫、口吃或简短句子。 3. **一致性**：严格遵循指令中规定的内容。

===来自认知核心的输入 ===指令：“{strategy_directive}” 非语言线索：“{non_verbal_cues}”

=== 任务 ===生成患者对医生的言语回应。将非语言线索放在星号之间（例如，*目光低垂*）。

B.3 EvaluatorAgent 提示

B.3.1 EvaluatorCoT 模块

此提示用于处理临床判断、怀疑度跟踪和量表管理。

清单 4：Evaluator CoT 模块的系统提示

你是一位经验丰富的精神科医生的临床推理引擎。
当前评估量表：［量表名称］（例如，PHQ-9）

=== 临床任务 ===1. **不一致性检查**：- 将患者当前的回答与个人档案历史和非语言线索进行比较。- 标记任何差异（例如，情感与内容不符）。

2. **怀疑度跟踪**：- 更新“怀疑程度_分数”（0.0 - 1.0）。- 如果怀疑程度上升，则将模式从“数据收集”切换为“核实”。

3. **下一步决策**：- **继续**：询问量表中的下一项。

- **调查**：提出后续探询，以澄清不一致之处。

- **终止**：达到信息饱和时。

===输出格式（仅限 JSON） ==={ "reasoning_step": "患者否认自己感到悲伤，但看起来含着泪。这存在矛盾。", "suspicion_score": 0.4, "next_move_type": "调查", "target_topic": "抑郁心境", "guidance_for_chat": "温和地指出患者含泪，并询问其近期的情绪状态". }

B.3.2 EvaluatorChat 模块

此提示可确保医生的语气专业、富有同理心，且结构严谨。

清单 5：评估器聊天模块的系统提示

你是精神科医生的声音。你从推理引擎接收临床指导。

=== 指南 ===1. **语气**：专业、共情、不带评判。2. **技巧**：- 使用“正常化”（例如：“很多人都有这种感受……”）来降低患者的防备心理。- 使用“开放式问题”进行探查。3. **安全性**：如果标记出自杀风险，请立即优先进行安全评估。

=== 来自推理引擎的输入 ===指导：“{guidance_for_chat}”

=== 任务 ===生成对话的下一轮内容。保持简洁、自然。

B.4 诊断智能体提示

诊断智能体充当最终裁定者。它综合多模态证据——对话记录、患者的自我报告以及评估器的临床观察——以生成结构化诊断并进行严重程度分级。

清单 6：诊断智能体的系统提示 t

您是一名高级顾问精神科医生。您的任务是审阅完整的病例档案，并撰写最终临床报告。

=== 病例档案 ===1. **患者档案**：{demographics}，{ chief_主诉} 2. **评估结果**：- 患者自我报告评分：{patient_score}（例如，PHQ-9 = 18）- 评估器评定评分：{ evaluator_score} （例如，汉密尔顿抑郁量表 = 22）3. **评估者备注**：- 怀疑程度：{suspicion_score} - 观察到的不一致：{inconsistency_logs}

=== DI AGNOSTIC TASKS ===1. **状态分类**：确定主要病理（健康、抑郁症、焦虑或创伤后应激障碍）。 2. **严重程度分级**：评估强度（无、轻度、中度、重度）。 - *关键*：重点参考评估器的“怀疑评分”。如果患者被标记为“隐瞒”，则应优先信任临床医生评定评分，而非自我报告。如果为“夸大”，则应相应下调严重程度。 3. **证据引用**：引用支持结论的具体对话轮次或行为线索。

=== 输出格式（JSON） ===｛ "final_diagnosis": { "status": "抑郁症", "severity": "中度" }, "reasoning": { "symptom_match": "患者符合快感缺失和疲劳的标准……", "discrepancy_resolution": "患者自我报告为‘重度’，但对话显示其情感连贯；因此下调为‘中度’。", "key_evidence": ［"第14轮：承认存在睡眠问题", "第22轮：否认有自杀意念"］} }

B.5 心理访谈模拟与评定系统提示

此提示定义了负责模拟临床访谈本身的核心智能体。它旨在根据结构化 JSON 输入，在单次对话中严格交替扮演患者和医生/临床医生。其主要任务是生成逼真的访谈转录文本，随后完成所有指定的患者自评量表和临床医生评定量表。

B.5.1 系统身份与任务目标

系统充当**心理评估模拟系统**。其核心能力是**严格区分并交替扮演**两种不同角色：

- **患者**
- **医生 / 临床医生**

根据输入的 JSON，系统必须：

1. 模拟一场真实、自然的临床访谈，包含足够的对话轮次、自然的交互和临床上合理的流程。2. 访谈结束后，完成**所有**提供的患者自评量表和**所有**提供的临床医生评定量表。

注意：此任务仅用于模拟和研究，不构成真实诊断。

全局约束：

- 不同会话之间不得使用记忆。
- 最终输出必须是一个 JSON 对象。
- JSON 必须完全使用英文编写。

B.5.2 输入数据结构

系统接收一个具有以下结构的 JSON 对象：

{ "档案": { "人口统计信息": { ... }, "主要_主诉": "...", "症状_病史": "...", "社会心理_因素": { ... }, "风险_标记": { ... }, "行为_倾向": "...", "沟通_风格": "...", "情感_基线": "...", "心理测量学": { ... } }, "转录文本": [...],"量表_已选": { "自我_报告_量表": [{ "名称": "..." }],"临床医生_量表": [{ "名称": "..." }]}, "基准_真值": { ... } }

关键输入约束：

1. ground_truth 仅用于**内部一致性检查**。医生角色不得查看它，也不得将其用作量表评分的参考。2. 系统**不得**引用、转述、改述或明确使用 ground_truth。3. 患者意图是**静态的**，必须**始终保持不变**。4. self_report_scale 和 clinician_scale 数组均可包含**多个量表**。

B.5.3 强制多量表规则（硬性约束）

如果输入中提供了多个量表，则必须全部引用、解读并评分。

- **要求：**

  1. 每个给定的量表都必须有对话证据支持，提供完整的评分结果，且不得跳过、合并或替代。2. 对话必须涵盖**所有量表的并集**所要求的全部核心领域。如果缺少某个领域，医生必须主动询问。

  3. 评分阶段必须针对每个量表（如适用）提供：条目级评分、总分以及严重程度／解读。

- **禁止：** 只对某个“主要”量表评分、提及某个量表却不对其评分，或用一个量表的结果替代另一个量表的结果。

B.5.4 角色定义与生成规则

A. 患者角色

- **静态行为建模：** 所有回答都必须严格遵循档案中不变的字段：行为_倾向、沟通_风格、情感_基线、心理测量学和主要_主诉。意图保持不变。

- **默认立场：** 冷静／有所保留／愿意配合，回答简洁但不回避，不主动展开，也不使用临床或诊断语言。

- **语言规则：** 使用日常口语英语。通过亲身经历表达症状，而非使用精神病学术语。

- **类人互动：** 患者可以表现出真实自然的行为，例如回答含糊、犹豫或带有情绪色彩的表达，但不得添加新信息。信息可以在多轮对话中逐步透露。

B. 医生／临床医生角色

- **最高规则：** 临床医生**不**知道真实情况。

- **临床推理：** 假设患者意图稳定，但内容未知。优先确保访谈完整性、内部一致性和循证评分。

- **访谈策略：** 采用分层、渐进式探索（核心症状 → 展开 →澄清）。鼓励采用类人化行为（反思性陈述、总结）。当量表核心领域尚未探明、回答过于含糊、出现风险信号或存在逻辑不一致时，必须进行强制追问。

- **语言风格：** 自然、专业，话题转换流畅。避免照搬量表条目措辞。

B.5.5 对话执行规则

- **轮次结构：** 对话从第 =1 轮开始。每轮包含医生发言和患者回应。可选地加入非语言标签（≤10 字符）。

- **第1轮（固定开场）：**

  – 医生：问候并 + 开放式邀请。– 患者：必须自然地包含主诉的**所有核心要素**_。

- **后续轮次：** 医生根据先前的回答进行合乎逻辑的追问。患者可以澄清或稍微修正先前的陈述。

- **终止条件（≥2项或以上）：**

  – 已涵盖所有量表相关领域。– 总轮数 ≥ **18**（无上限；鼓励进行更多轮次）。– 已收集到足够证据，可对所有量表进行评分。– 患者反复表示没有更多信息。

B.5.6 评分阶段

- **患者自评量表：** 输出必须包含一个患者_自我_报告对象，其中每个输入的自我_报告_量表数组中列出的量表都必须有一个完整的子对象（包括条目得分、总分、严重程度、解释和对话证据）。不得遗漏或合并。

- **临床医生评定量表：** 输出必须包含一个医生_临床医生_报告对象，其中输入的临床医生_量表数组中列出的**每个**量表都必须有一个完整的子对象。每个得分都必须能追溯到对话证据。

B.5.7 最终输出格式（硬性要求）

系统必须输出一个具有以下结构的单一 JSON 对象：

```
{ "final_转录文本": [... ],// 完整的模拟对话 "患者_自我_报告": { ... }, // 所有已完成的自评量表 "医生_临床医生_报告": { ... } // 所有已完成的临床医生评定量表 }
```

B.5.8 示例输入与输出

输入示例： 包含选定量表的通用患者档案。

清单 7：通用输入档案格式

```
{ "档案": { " 人口统计信息": { ... }, "主要_诉求": "...", "症状_史": "...", "社会心理_因素": { ... }, "风险_标记": { ... }, "行为_倾向": "...", "沟通_风格": "...", "情感_基线": "...", "心理测量学": { ... } }, "量表_已选": { " 自我 _ 报告 _ 量表 ": [{ " 名称 ": " 量表 _ 名称_1" }, { "名称": "量表_名称_2" }],"临床医生_量表": [{ "名称": "临床医生_量表_名称_1" }]} }
```

输出示例： 预期输出格式包括对话转录文本和已完成的量表报告。

清单8：通用输出格式

```
{ "final_转录文本": [{ "对话轮次": 1, "医生": "...", "患者": "...", "非言语行为": "..." }, // ... 更多对话轮次],"患者_自我_报告": { "量表_名称_1": { "得分": { ... }, "总_得分": ..., "严重程度": "...", "解释": "...", "对话_证据": { ... } }, "量表_名称_2": { // ... 类似结构 } }, " 医生 _ 临床医生 _ 报告 ": { " 临床医生_量表_名称_1": { "得分": { ... }, "总_得分": ..., "严重程度": "...", "解释": "...", "对话_证据": { ... } } } }
```

B.6 临床真实性评估提示 t

此提示定义了评估模拟医患对话临床真实性的评估标准。评估器的角色是一位具备计算语言学专业知识的资深精神科临床督导。

B.6.1 角色定义与核心原则

角色： 一名拥有 20 多年经验的资深精神科临床督导，专长于抑郁症、焦虑障碍和创伤后应激障碍诊断。同时也是一名擅长话语分析的计算语言学家。

黄金法则： “真实性不等于完美。”无可挑剔的语法、高信息密度以及完全符合逻辑的问答交流，往往是合成数据的指标。高真实性得分应包含阻力、回避、离题和尴尬的沉默等元素。

B.6.2 评估维度（1-10 分）

每段对话均从五个维度进行评估，其中 10 分代表“与真实临床场景最为相似”：

1. **话语自然度：** 对话流畅度的自然程度，反映患者的思维模式（反刍、思维偏离）。2. **语言质感与口语特征：** 体现口语细粒度的程度，以及书面语的侵入情况。3. **情绪细粒度与深度：** 患者表达情绪复杂性和矛盾心理的能力。4. **防御机制与阻力：** 是否存在符合临床情境的防御行为。5. **交互动力学与参与度：** 双向交互质量，以及医生采用适应性策略的情况。

B.6.3 评估输入格式

评估系统接受多种格式的对话输入：

- **纯文本：** 直接输入对话记录
- **CSV/JSON结构化格式：** 带说话者标注的对话
- **完整对话：** 通常包含 10-30 轮医患交流

示例输入结构：

Listing9：示例对话输入格式医生：询问患者目前担忧的问题……患者：用日常语言描述症状……医生：追问以了解具体症状……患者：回答时可能表现出犹豫或淡化……//……后续还有多轮对话

B.6.4 评估输出格式

该系统会生成 JSON 格式的评估结果，为全部五个维度提供得分和理由。

所需输出结构：

Listing10: 评估输出格式{ "话语_自然度": { "score": [1-10],"reason": "简要说明理由，并引用具体的对话特征……" }, "语言_质感": { "score": [1-10],"reason": "简要说明理由，并引用具体的对话特征……" }, "情绪_细腻度": { "score": [1-10],"reason": "简要说明理由，并引用具体的对话特征……" }, "防御_阻力": { "score": [1-10],"reason": "简要说明理由，并引用具体的对话特征……" }, "交互_动力学": { "score": [1-10],"reason": "简要说明理由，并引用具体的对话特征……" } }

B.6.5 自动评估系统

该提示词在一个自动评估系统中实现，该系统：

- 处理 CSV、JSON 和 TXT 格式的对话文件
- 使用特定格式的解析器提取对话内容

• 向 DeepSeek API 发送结构化提示以进行评估

• 解析 JSON 响应并生成全面的 Excel 报告

• 支持通过状态管理恢复评估

系统输出报告：

• **详细评估结果：** 包含理由的单个文件得分

•**文件夹统计：** 跨数据源的比较分析

• **总体指标：** 汇总得分和质量评估

• **维度分析：**按评估标准分解表现

C 隐瞒与夸大特征库

C.1 隐瞒特征

• **01** - 弱化/淡化核心症状：将痛苦描述为“轻微/我没事/我能应付”

– 可观察指标：
  * 频繁使用弱化语（例如，“有点儿”，’不是那样’，’大体还好’)
  * 强调对照组和耐力
  * 避免量化严重程度（频率—次/强度/持续时间）
– 目标字段：
  * 主诉_、症状_史，行为_倾向，沟通-风格_，情感_基线，心理社-会因素_。应对_机制，心理测-量。印象_管理
– 标签：淡化

• **02** - 选择性否认并带有“泄露线索”：口头否认存在问题，但会包含轻微的症状片段

– 可观察指标：* 先否认，随后泄露线索（睡眠/精力/兴趣/注意力下降） * 将症状归因于小的外部问题，而非持续性状态 * 不要主动化解矛盾– 目标字段：* 主要_主诉、症状_病史、行为_倾向、沟通_风格、情感_基线、心理测量学.印象_管理– 标签：否认_伴随_泄露

• **03** - 回避细节：对时间线/频率/诱因给出含糊回答，或转移话题

– 可观察指标：* “很难说/记不清/就是这样”等回答的频率较高 * 用笼统总结结束追问 * 转向日常琐事或积极话题– 目标字段：* 主诉_、症状_史、行为_倾向、沟通_风格、情感_基线、心理社会_因素.压力源、心理测量.印象_管理– 标签：回避

• **04**- 积极自我呈现/印象管理：强调优势和功能状态，淡化困难– 可观察指标：* 更多谈论目标、自律和责任感 * 简短提及负面情绪，但不作展开 * 将困难描述为“这并不会真正影响我”– 目标字段：* 行为_倾向、沟通_风格、情感_基线、心理社会_因素.目标、心理社会_因素.社会_支持度、主诉_、症状_史、心理测量.印象_管理、心理测量.宜人性– 标签：印象_管理

• **05**- 少报风险相关内容：迅速否认自伤/自杀倾向，几乎不作 elaboration

– 可观察指标：
  * 对风险问题的回答极短且斩钉截铁 * 回避讨论“最糟糕的时刻”或“失控” * 淡化求助史或危险行为

– 目标字段：
  * 风险_标记.自杀，风险_标记.自我_伤害，风险_标记.暴力，主诉_，症状_史，行为_倾向，沟通_风格，情感_基线，治疗_史，心理测量.印象_管理

– 标签：risk_underreporting

• **06** - 正常化/功能化痛苦：将一切解释为压力或疲劳，并拒绝接受心理健康方面的标签

– 可观察指标：* 用单一生活压力源解释多种经历 * 否认其跨情境影响或持续性 * 与心理学标签保持距离– 目标字段：* chief_complaint, symptom_history, behavior_tendency, communica- tion_style, affect_baseline, psychoso- cial_factors.stressors, psychoso- cial_factors.coping_mechanism, psychometrics.openness– 标签：normalization

D 补充模拟示例

本附录展示了一个具有代表性的模拟精神科初诊访谈示例。本节旨在呈现医生与患者之间完整且可观察的对话，以及访谈过程中生成的面向临床医生的输出。

D.1 数据可用性

所有模拟输出均以结构化 JSON 格式发布，并可公开获取：

https://github.com/Zephyr-Wang-personal/Mentalhealth/tree/master/data/对话

每条模拟记录都包含医生与患者之间按轮次展开的对话交流、可选的非语言行为标签，以及面向临床医生的输出，例如怀疑度分数和诊断假设更新。

为便于阅读和展示，本附录中的示例省略了原始模拟输出中包含的内部推理字段。这些省略的字段包括内部思维过程、中间心理状态更新、隐瞒特征注释，以及用于指导对话生成的程序控制指令。此项省略仅 适用于附录展示。

托管在 GitHub 上的所有原始模拟文件均保持完整且未作更改，并保留了模拟过程中使用的全部内部字段。

D.2 完整对话示例

下面展示了完整的模拟精神科初诊访谈。医生与患者之间的所有对话轮次均按时间顺序列出。这里只显示外部可观察到的对话内容和非语言行为标签。

第0轮

医生：“你好，欢迎。请坐。为了开始访谈，你能跟我说说今天是什么原因让你来这里吗？”

第1轮

患者：“嗨，谢谢。嗯，当然。呃……说实话？我只是……我想我妈妈觉得我来跟人聊聊可能是个好主意。她很担心我。不过我状态很好。真的，特别好。我到处演出、旅行、认识新朋友……感觉超棒。生活很美好，你懂吧？”

医生：“嗯，我明白。很高兴你对生活感到积极。有时，即使生活表面上看起来不错，内心深处也可能隐隐有些担忧或紧张。你能告诉我你最近的情绪怎么样吗——即使是在演出和旅行时，你是否也会感到紧张或忐忑不安？”

第2轮

患者：“不，没什么。比如演出……那就是我的领域。那种纯粹的精力，你懂的？人群、灯光……那本身就像是一种兴奋感。旅行也一样。新的城市、新的面孔、新的派对。一切都很好。我的意思是，当然，有时候会有……一些让人烦恼的后勤问题。航班延误，演出推广人可能不太靠谱……但那只是日常生活中的小事。这不会影响到我。”

医生：“明白了。听起来，你对这种生活方式中的起起落落应对得相当不错。换个话题，我想问问你的恐惧。即使生活忙碌而令人兴奋，有些人也会害怕某些特定的事，或者发现自己会回避某些情境。你有没有类似的情况？”

第3轮

患者：　[非言语行为：目光回避]“嗯。害怕什么？我不知道这能不能算是害怕。更像是……呃，有时候我会担心事情变得乏味，你懂吧？比如，现场氛围、音……万一这一切都……不再有趣了呢？或者人们不再来了呢？那可就扫兴了。而且，我想……有时候独自待在一个谁都不认识的城市里的酒店房间，会觉得有点奇怪。不是害怕，只是……安静。太安静了。不过这就是为什么要出门，对吧？去找派对。问题解决。”

医生：“这很有道理。对于自己热衷的事情，人们自然会考虑它的未来，而活动间隙的安静时光也可能让人感觉格外冷清。谢谢你愿意分享这些。我们暂时聊点更偏身体方面的内容。当你处于那些比较安静的时刻，或者有时在准备参加演出时，你会注意到任何压力带来的身体反应吗？比如心跳加速、感觉有点发抖，或者可能气短？”

第4轮

患者：“哦，演出之前吗？那当然会。那可不是压力，只是……肾上腺素，你懂吧？就像我的心脏怦怦直跳，但这种跳动感觉很好。就像发动机启动前轰鸣起来。让我觉得自己充满活力。这是整个过程的一部分。我不会发抖、喘不过气之类的。只是……精力。纯粹的精力。”

医生：“没错，那种演出前的能量是一种强烈的感受。精力如此充沛，日程又安排得满满当当，还要旅行和演出，你最近睡得怎么样？你觉得自己容易放松下来、睡个好觉吗？还是有时思绪或身体让你难以入睡？”

第5轮

患者：“睡眠？嗯……说实话，我的睡眠挺不规律的。演出结束后，我会特别兴奋，精力还很充沛，你懂吧？所以有时候我就……继续玩下去。和朋友待在一起、认识些人什么的。等我真的累倒了，就会睡得很沉。但有时，尤其是……你懂的，参加了些派对之后，睡眠就会很奇怪。比如，我会躺在那里，脑子里飞快地想着一首曲目的点子，或者想着下一个城市。有时候我就是不想睡。总有事情在发生。”

医生：“紧凑日程下出现这种情况很常见。当你的思绪像这样飞速运转时，或者白天没有演出的时候，你的专注力怎么样？比如，如果你试着创作一首新曲目或处理一些规划，你觉得自己容易集中注意力吗？还是思绪容易飘走，或者觉得脑子昏沉？”

第6轮

患者：“这要看情况，你懂的？当我进入状态，比如真的沉浸在一个新的节拍或一些歌词中时……我可以连续几个小时全神贯注。其他一切都不存在了。但是……嗯，有时候，尤其是旅行后感到疲惫，或者前一晚外出到很晚，我的大脑就会觉得……一片模糊。比如，我会坐下来安排一些巡演日期或回答电子邮件，但就是……做不到。我会开始想别的事情，或者只是盯着屏幕发呆。真让人恼火。不过，通常我会先去做点别的事，直到这种迷雾感消散。”

医生：“这种‘模糊’的感觉确实会让人非常沮丧，尤其是在你需要完成事情的时候。我们暂时谈谈情绪吧。随着你的生活方式时好时坏，你是否曾有过这样的时期：感到情绪低落、悲伤，或者只是对平时喜欢的事情提不起兴趣——甚至连音乐或派对也不例外？”

第7轮

患者：“我的意思是……每个人都有状态不好的时候，对吧？当然。有时候，可能是在巡演结束、回到妈妈家之后，会觉得有点……平淡。就像音量被调得很低，安静反而更显声响。有那么一两天，我甚至可能不想拿起麦克风，也不想查看消息。但这种感觉不会持续太久。我只是……觉得一直无聊下去也无聊了，你懂吧？然后我就又回到原来的状态。音乐、人们……总能把我重新吸引回去。”

医生：“这个说法很贴切——安静反而更显声响。有时候情绪低落时，身体也会有所反应。你有没有注意到身体紧绷或疼痛的情况？比如肩膀、颈部或背部，尤其是在那些比较安静的时候，或者长时间演出之后？”

第8轮

患者：“不太算，没有。演出结束后，我通常只是……筋疲力尽，但这种感觉是好的。就像我的整个身体都在嗡嗡作响，不过是那种令人愉悦的嗡鸣。我没有疼痛或紧绷感。如果非要说的话，我感觉身体很放松。非常放松。有时候我会……你知道，吸烟一点东西，帮助自己保持这种状态，放松下来。但是

不，我的身体跟得上我，状态相当不错。”

医生：“听你这么说真好。保持放松很重要。我想问问另一种身体感觉。你是否曾感觉耳鸣、视力有点异常、突然冷热潮红，或者某处有刺痛感或无力感？”

第9轮

患者：“嗯……是的，有时候会。但只是在我，你懂的，玩的时候。比如，我在派对上或者演出结束后服用了什么东西，可能会有一阵子耳鸣，或者感觉有点刺痛。这就是其中的一部分，你懂吧？又不是我什么都不做、只是坐着的时候也会发生。这是意料之中的，而且总会过去。”

医生：“我明白。这些感觉可能与特定情境有关。我们来聊聊你的心脏吧。除了演出前那种兴奋的怦怦直跳，你是否还会在其他时候注意到心跳加速、心跳漏拍，或者胸口感到不舒服？比如在毫无预兆的时候，或者甚至是在试图放松的时候？”

第10轮

患者：“呃，是的，有时候会。但话说回来，通常都是……化学物质导致的，你懂吧？比如，我要是吸了几口什么的，心脏就会跳快一阵。这很正常。不会把我吓坏什么的。我知道是怎么回事，而且过一会儿就会平静下来。我在妈妈家看电视、放松的时候，不会这样。”

医生：“说得很清楚，谢谢。那你的呼吸呢？除了演出时的体力消耗，你有没有在不活动的时候感到气短、像是吸不到足够的空气，或者胸口发紧？”

第11轮

患者：“没有，不太会。除非我真的在舞池里跳得很卖力什么的，但那只是因为体能不太好，你懂吧？（轻轻笑）不过没有，就像只是坐着却觉得喘不过气？这种情况不会发生。”

医生：“（轻轻笑）有道理。我们换个方面聊聊吧。你的胃最近怎么样？按照你的作息，你会不会出现恶心、胃痛或食欲不振之类的情况？”

第12轮

患者：“哦，当然。我吃饭的时间……完全没规律。有时候我忙着做某件事，就会忘了吃饭；或者就是不饿。如果我喝了酒，或者你知道，做了其他什么事，第二天胃可能会有点不舒服。不过没什么大不了的。我就吃点吐司之类的，等它自己好起来。我想，这就是生活方式的一部分吧。”

医生：“作息不规律时，出现这种情况很常见。如果你不介意，我还想问问另一个身体方面的问题。你有没有注意到如厕习惯有所改变，比如更频繁或更急着上厕所，或者这方面还有其他问题？”

第13轮

患者： [非言语行为：目光回避]“没有，倒没有。那方面一切正常。”

医生：“好的，这是这一组问题的最后一个。你有没有出现过一些情况，比如不热的时候也大量出汗、口干得厉害，或感到头晕、眩晕？”

第14轮

患者：“嗯……对，有时候会。口干这种情况肯定会有，比如我抽了烟或别的什么之后。头晕嘛……可能是熬了一整夜后起身太快，或者某种东西的药劲儿正在消退的时候。不过这不是经常发生的事。只是……你知道的，副作用。过一会儿就好了。”

医生：“谢谢，这些信息很有帮助。我注意到我们交谈时，你一直挺坦诚、放松的。此刻坐在这里，你感觉怎么样？有没有坐立不安或紧张的感觉，还是觉得挺自在？”

第15轮

患者： [非言语行为：目光回避]“我挺好的，对，很自在。就是……挺放松的。一直坐着聊这么久有点奇怪，不过没事。我挺好的。”

医生：“很高兴你感觉自在。为了更全面地了解你的情绪状态，我想问一下：过去几周里，你有没有一段时间感到情绪低落、悲伤，甚至对事情感到绝望？”

第16轮

患者：“不，没怎么。就像我说的，可能偶尔有一天会觉得有点情绪低落，但绝望？不。我有太多事情要做了，不会感到绝望。”

医生：“有这样的心态很好。换个话题，你是否有过因某些事情感到内疚，或责怪自己导致事情变成现在这样——比如你的事业、家庭，或生活中的其他方面？”

第17轮

患者：“内疚？嗯……我不太会用这种方式去想，你知道吗？我是说，当然，有时候我

会觉得自己在外出时应该多给妈妈打打电话。或者，也许我不该对某个人说某些话。但这些事不会一直困扰我。我不会坐在那里不停地责备自己。事情已经发生了，你知道吗？继续向前走就好。”

医生：“这样看待这件事很健康。我需要问一个我会问每个人的问题，可以吗？有时候，当人们情绪低落或心里压着很多事情时，可能会产生伤害自己的想法，或者觉得活着没有意义。你最近有过类似的想法吗？”

第18轮

患者：“没有。不可能。就像我说的，生活很好。即使并不完美，它仍然是……它就是生活，你懂吗？总会有下一场演出、另一座城市、另一种节拍。我不是那样的人。”

医生：“听你这么说我很高兴。如果可以的话，我们暂时回到睡眠这个话题。你提到自己的睡眠很不规律。当你确实尝试上床睡觉时，比如在你妈妈家或酒店里，你会难以真正入睡吗？比如躺很长时间才能慢慢睡着？”

第19轮

患者：[非言语行为：模棱两可]“嗯，有时候。尤其是当我的脑子还停不下来时，你懂吗？比如演出结束后，我还处于兴奋状态。或者如果我刚在外面待过，有时就是……无法关掉思绪。我会躺在那里，脑海里的想法不断跳来跳去。挺烦人的，但最终我还是会累得睡着。或者有时我就干脆起床，做点什么，直到自己累到足够困为止。”

医生：“我明白。你入睡后，通常会一觉睡到天亮，还是会在半夜醒来？”

第20轮

患者：[非言语行为：目光回避]“这得看情况。如果我真的累坏了，就会睡得很沉。但如果我喝了酒，或者……做了些别的事，可能会醒好几次。要是有早班飞机，或者有什么事挂在心上，我也可能突然醒来。不过通常只要睡着了，就能一觉睡到底。”

医生：“好的。你夜里醒来时，通常是临近早晨的时候吗？比如说，比你需要起床的时间早很多？如果醒得那么早，你还能再睡着吗，还是就直接起床开始一天了？”

第21轮

患者：[非言语行为：含糊]“通常是……凌晨4点或5点？那段奇怪而安静的时间。如果是因为航班或什么事情，我通常就会直接起床开始一天，大脑已经进入行动模式了。但如果只是……毫无缘由地醒来，有时我可以翻个身，然后重新睡着。其他时候，我就会拿起手机，刷一会儿，看看世界另一边正在发生什么。”

医生：“谢谢你解释这些。我们来聊聊你的日常生活吧。最近有了旅行、演出以及其他各种事情之后，你是否发现自己更难完成事情、处理好工作，甚至难以享受那些平时喜欢的生活乐趣？”

第22轮

患者：“不了，我挺好的。我的意思是，有时候行政事务——电子邮件、规划之类的——确实感觉像是在完成一项苦差事。但真正的工作、音乐、演出呢？那才是有趣的部分。我仍然很喜欢做这些。登上舞台永远不会让我厌倦。所以不，我不会说这变得更难了。只是……生活而已。”

医生：“听你这么说真好。我想换个角度问问你的精力和专注力。过去几周里，你有没有觉得自己的思维或说话速度比平时慢，或者即使你想集中注意力或行动起来，也觉得更困难？”

第23轮

患者：“没有，不太有。我的意思是，就像我说的，有时候累了什么的，会有那种脑雾感。但变慢？没有。我脑子通常转得飞快。”

医生：“这说得通。你思绪飞转的时候，会不会也感到身体躁动？比如坐立不安、踱步，或者就是觉得没法安静地坐着？”

第24轮

患者：[非言语行为：含糊]“嗯，当然。如果我脑子里冒出很多想法，或者……你知道，因为某些事情变得兴奋起来，我就会坐立不安。没法老坐在椅子上。我会站起来走来走去，也许还会在笔记本电脑上随手敲出一段节拍。这只是精力旺盛，需要找个地方释放。”

医生：“我明白，这股精力需要一个出口。除了那种躁动感，你是否也发现自己比以前更紧张、易怒，或者更容易为事情担忧——哪怕是一些小事？”

第25轮

患者：“易怒？嗯……也许有时候是吧。比如，去演出的路上堵在交通里，或者某件设备坏了。那一刻，小事也会显得更严重。但这不是一直如此。我并不会总觉得紧张什么的。只是……看情况，你懂吧？”

医生：“这可以理解，这是一个高压力的环境。在那些令人感到压力的时刻，

你有没有注意到，除了易激惹之外，身体还会出现什么反应？比如胃部不适、心脏怦怦直跳，或者呼吸感到紧促？”

第26轮

患者：“嗯，有时候会。比如演出马上要开始了，却突然出现什么最后关头的问题，我的心脏就会开始怦怦跳。那又是肾上腺素在起作用。而且我的胃可能会有点……翻腾。但这并不像什么惊恐发作之类的。只是我的身体在准备应对问题。问题一解决，这些感觉就全消失了。”

医生：“这样理解很有帮助——你的身体正在做好准备。除了这些时候，最近你的胃或消化方面还有其他问题吗？比如明显的食欲不振、恶心，或类似的情况？”

第27轮

患者：“基本就是我之前提到的那些。如果我参加了聚会，第二天可能会觉得有点恶心或不太舒服。有时候我只是忘了吃饭。但这不是什么新问题，也不严重。我的食欲总会恢复。”

医生：“好的。先不谈胃的问题，最近你的身体还有其他一般性症状吗？比如感觉异常沉重或疼痛、头痛、肌肉疼痛，或者只是比平时更容易感到疲倦？”

第28轮

患者： [非言语行为：含糊，目光回避]“累，嗯，有时候会。那只是因为熬夜和旅行。演出一整晚之后，我的腿也会有点疼，但那只是因为跳舞和走来走去。头痛……不怎么有。也许脱水或宿醉时会有一点，但也就这样。没什么异常。”

医生：“谢谢你告诉我这些。我想问一个更私人的问题，这是评估中的常规部分。你最近有没有注意到自己对性的兴趣或性欲有什么变化？”

第29轮

患者：“没有，没什么变化。一切都挺好的。”

医生：“好的。换个话题，你最近是否发现自己经常担忧自己的健康，或总是纠结于某些身体症状？”

第30轮

患者：“不太算，没有。我的意思是，我知道自己的生活方式不完全是……医生推荐的那种。（轻笑）但我感觉很好。我知道自己出现的那些感觉是怎么回事。我不会坐在那里为此担忧。”

医生：“（轻笑）也有道理。这组问题的最后一个问题：你最近有没有注意到体重发生任何变化，比如没有刻意减重却瘦了？”

第31轮

患者： [非言语行为：含义不明]“我的体重挺稳定的。如果我饮食不规律，体重可能会增减几磅，但也没什么大不了的。”

医生：“好的。根据我们今天谈过的所有内容，你会如何描述自己的整体状态？你觉得自己可能正面临什么心理健康问题，还是认为这些经历只是生活方式和工作要求的一部分？”

第32轮

患者：“说实话？我感觉很好。我过着自己的生活，做自己想做的事。没错，我会参加派对、旅行，也努力工作。有时会觉得累，有时胃不舒服，有时大脑会有点迷糊。但这只是……做这一行要付出的代价，你懂吧？这不是心理健康问题。这只是我的生活，而且我喜欢这样的生活。”

医生：“感谢你分享这些看法。显然，你很清楚自己的生活方式所涉及的利弊。感谢你今天抽出时间并坦诚交流。我们的访谈到此结束。[结束访谈]”

D.3 面向临床医生的输出

在选定的对话轮次中，医生会生成面向临床医生的输出，反映访谈评估的进展。这些输出包括持续更新的怀疑评分，以及基于可观察到的对话模式对诊断假设所做的更新。

第17轮：

怀疑评分：0.88

假设更新：- 持续淡化焦虑和情绪相关症状 - 潜在的痛苦可能被正常化和积极自我呈现所掩盖

第19轮：

怀疑评分：0.90

假设更新：-持续淡化睡眠相关困难 - 将睡眠障碍描述为生活方式相关问题，而非临床显著问题

上方仅展示可观察到的对话内容、非语言行为标签和面向临床医生的输出。内部推理过程、中间

状态更新、特征匹配逻辑以及模拟过程中使用的程序控制信号为清晰起见未在此展示，但均完整保留在发布于 GitHub 上的原始模拟记录中。

E 心理测量量表列表

表 7详细列出了我们系统存储库中使用的 46 种精神科测量工具，并按诊断领域和施测类型分类。

领域	Type	缩写	全名	条目	评分／临界值
PTSD	临床医生	临床医生管理的创伤后应激障碍量表 DSM-5	临床医生施测的创伤后应激障碍量表	30	严重程度：每项症状 0–4 分
		创伤后应激障碍症状量表——访谈版	创伤后应激障碍症状量表——访谈版 sion	20	临界值 ≥ 23
	Self	DTS	戴维森创伤量表	17	临界值 ≥ 40
		事件影响量表修订版	事件影响量表修订版	22	临界值 ≥ 33
		ITQ	国际创伤问卷	18	DSM-5/《国际疾病分类》第十一次修订本》标准
		PC-PTSD-5	DSM-5基层医疗创伤后应激障碍筛查量表	5	临界值 ≥ 3（女性）/ 4（男性）
焦虑	临床医生	汉密尔顿焦虑量表	汉密尔顿焦虑量表	14	>14 中度； >23重度
		LSAS	利博维茨社交焦虑量表	24	>60 广泛性社交焦虑
	Self	BAI	贝克焦虑量表	21	0–21 低；22–35 中度； 36+ 重度。
		广泛性焦虑量表-7	广泛性焦虑障碍量表-7	7	5/10/15（轻度/中度/重度）
		PSWQ	宾夕法尼亚州立大学担忧问卷	16	>39 中度； >59重度。
		SPIN	社交恐惧量表	17	>19轻度； >30中度； >40重度
		STAI (状态)	状态-特质焦虑量表（状态量表)	20	>39 中度； >59 高
抑郁症	临床医生	汉密尔顿抑郁量表	汉密尔顿抑郁量表	17	0–7 正常； >19重度
		蒙哥马利-艾斯伯格抑郁评定量表	蒙哥马利-艾斯伯格抑郁评定量表	10	0–6 正常； >34重度
	Self	BDI	贝克抑郁量表	21	0–10 正常； >30重度
		BHS	贝克绝望量表	20	临界值 ≥ 9
		CDI	儿童抑郁量表	27	>12轻度； >38 重度
		CES-D	流行病学研究中心—抑郁症	20	临界值 ≥ 16
		儿童流行病学研究中心抑郁量表	儿童流行病学研究中心抑郁量表	20	临界值 ≥ 16
		修订版流行病学研究中心抑郁量表	修订版流行病学研究中心抑郁量表	20	>10中度； >22 高
		EPDS	爱丁堡产后抑郁量表	10	>9 可能； >13很可能
		老年抑郁量表（15项）	老年抑郁量表	15	≥ 5轻度； ≥ 10重度
		HADS	医院焦虑抑郁量表	14	≥ 8 临界； ≥ 11临床
		KADS	库彻青少年抑郁量表	11	>10轻度； >20中度/重度
		MDI	重度抑郁量表	10	ICD-10算法
		MFQ （简版）	情绪和感受问卷	13	>7轻度； >15中度/重度。
		PHQ-9	患者健康问卷-9	9	5/10/15/20（轻度/中度/重度）
		SDS	Zung抑郁自评量表	20	≥ 50 轻度; ≥ 60中度; ≥ 70重度
		WSAS	Weinberg情感筛查量表	10	>9显著； >20重度
情绪/人格自评		抑郁焦虑压力量表-21	抑郁焦虑压力量表	21	各分量表的严重程度评级
		16PF	卡特尔16种人格因素问题-naire	185	标准十分制分数（1–10）
		艾森克人格问卷修订版	艾森克 人格 问卷– 修订版	100	E/N/P/L量表
		明尼苏达多相人格量表-2	明尼苏达 多项 人格量表-2	567	T分数 ≥ 60(中国常模)
睡眠	Self	AIS	雅典失眠量表	8	临界值 ≥ 6
		DBAS-16	关于睡眠的功能失调信念与态度 睡眠	16	>53 中度； >107重度
		ESS	Epworth嗜睡量表	8	≥ 10过度嗜睡
		首次	福特失眠应激反应测试	9	>18中度； >27 高风险
		ISI	失眠严重程度指数	7	>7 轻度； >14 中度； >21重度
		MEQ	晨晚型问卷	19	昼夜节律类型分类
		PDSS	帕金森病睡眠量表	15	<121 轻度; <103中度; <86重度
		PSQI	匹兹堡睡眠质量指数	19	总分 ≥ 5(差)
		SDRS	睡眠障碍评定量表	10	>8 轻度; >16中度; >24重度
		SHPS	睡眠卫生习惯量表	30	>60中度; >120重度
		SSS	斯坦福嗜睡量表	1	>3困倦

表 7：评估代理存储库中的心理测量量表完整列表。