arXiv:2611.32044v1 [cs.HC] 24 Nov 2026

CHORAL：基于患者生成的健康数据进行纵向预后推理的多智能体脚手架

YOUNGJI KOH，KAIST，韩国ESTHER KANG，KAIST，韩国GYUNA KIM，KAIST，韩国SEUNGWAN JIN，KAIST，韩国HYUNSOO LEE，KAIST，韩国KYUNGSIK HAN，韩国汉阳大学UICHIN LEE*，KAIST，韩国

纵向患者生成的健康数据可为预后推理提供依据，帮助了解患者的病情在随访就诊之间可能如何演变。现有方法会总结病程轨迹或预测症状状态，但在考察异质证据如何影响预后方面支持有限。我们提出 CHORAL，这是一种多智能体脚手架，先进行领域特定评估，再开展跨领域综合与审查，同时保留证据和修订记录。在一项由 12 名精神科医生参与的形成性研究启发下，我们对 CHORAL 进行了定量评估，并邀请 22 名临床医生参与评估。CHORAL 的平衡准确率与持久性基线相当（0.673 对 0.672），同时识别出持久性基线未能发现的从低到高的恶化病例中的 40%。与单智能体基线相比，其平衡准确率从 0.639 提高至 0.673，证据覆盖率从 70.1% 提高至 89.2%。在证据评估、替代性解释、提出质疑的依据，以及区分哪些内容可采信、哪些内容需要核实等方面，临床医生对 CHORAL 的评分显著更高。这些发现表明，多智能体审议能够通过揭示替代性解释和核实依据，为审查提供支持。

CCS概念： • **以人为中心的计算** → **交互系统与工具**。

其他关键词和短语：心理健康、患者生成的健康数据、数字表型分析、预后推理、多智能体系统、人机协作

1 引言

心理健康门诊护理通常涉及多次随访就诊，在此期间，临床医生会评估患者病情自上次就诊以来发生了怎样的变化，并为后续护理做好准备 [68]。为此，临床医生不仅会考虑患者的当前状态（诊断），还会判断近期临床病程是在恶化、改善、持续还是波动，并预测未来可能的临床病程（预后）。这种预后推理 是一种面向未来的推断，将关于患者当前状态的诊断推理与治疗

推理以及后续干预联系起来 [23, 49]。

然而，门诊随访通常间隔数周，这使得在两次就诊之间持续观察患者的日常行为轨迹变得困难 [25]。患者生成的健康数据（PGHD）可以通过纵向记录诊所以外的行为来填补这一空白，例如身体活动、睡眠、活动能力和智能手机使用 [30, 45, 64]。在心理健康护理中，这些行为变化可以为评估未来可能出现的、与抑郁、焦虑、压力和睡眠问题相关的担忧提供轨迹层面的证据

*通讯作者

作者联系方式：YOUNGJI KOH，KAIST，韩国大田，youngji@kaist.ac.kr；ESTHER KANG，KAIST，韩国大田，esther.kang@kaist.ac.kr；GYUNA KIM，KAIST，韩国大田，gyuna.kim@kaist.ac.kr；SEUNGWAN JIN，KAIST，韩国大田，seungwan@kaist.ac.kr；HYUNSOO LEE，KAIST，韩国大田，hslee90@kaist.ac.kr；KYUNGSIK HAN，汉阳大学，韩国首尔，kyungsikhan@hanyang.ac.kr；Uichin Lee，KAIST，韩国大田，uclee@kaist.ac.kr。

。然而，使用纵向患者生成健康数据进行预后推理仍然面临挑战。

首先，基于异构PGHD构建有意义的纵向轨迹颇具挑战。PGHD由多种行为信号构成，这些信号的尺度和时间模式各不相同，临床医生必须区分短暂波动与持续变化。这需要结合每位患者自身的基线，考量多个信号的幅度、持久性和变化方向 [18]。其次，在抑郁、焦虑、压力和睡眠问题等与心理健康相关的不同方面，观察到的行为轨迹可能具有不同含义。同一种行为变化可能与多个领域相关，因此难以从单一的领域特定视角理解患者病情 [12]。因此，预后推理需要在多个领域中考察相同的纵向证据，并考虑由此得出的领域特定预后评估之间的关联，而不是将轨迹归结为单一解读。第三，单靠PGHD无法充分解释行为变化的临床背景。要评估观察到的行为变化的意义，需要考量其在不同领域中的相关性，以及它对日常功能可能产生的影响 [4, 66]。然而，充分解读这类变化所需的情境信息未必包含在PGHD中。PGHD可能会显示发生了什么变化，但不会说明变化为何发生，或患者对此有何感受。因此，系统应区分哪些内容能够根据现有证据加以解读，以及哪些信息缺口需要在随访期间通过补充临床背景来填补。

HCI 和心理健康感知领域的既有工作通过 PGHD摘要、可视化和预测，部分应对了这些挑战。摘要和可视化将纵向行为信号加以组织，以帮助临床医生了解发生了什么 [75, 83], ，但对于考察所观察到的变化如何在不同领域引发不同的预后关注点，支持仍然有限。预测方法根据 PGHD 估计未来结局或风险 [17, 36], ，但通常侧重于最终预测，而不是纵向证据如何在不同领域之间得到解读。临床多智能体系统还将推理分配给专业化角色或阶段 [35, 62, 71], ，但对于在跨领域整合之前保留对同一纵向证据的独立评估，支持仍然有限。因此，在纵向 PGHD 上搭建多领域预后推理过程本身仍存在空白。

为了搭建基于 PGHD 的预后推理，临床医生不仅应能够查看系统的最终输出，还应能够查看支撑该输出的证据和推理。当系统的解读与他们自己的评估不一致时，临床医生应能够检查其理由，并在必要时质疑或反驳该判断[2, 41]。因此，中间推理应保持可检查性，包括支持领域特定的预后评估的证据，以及这些评估如何在不同领域之间整合。

基于这一动机，我们开发了**CHORAL**（**C**onte**x**t-aware **H**ea**l**th-data **O**r**c**hestration of **R**easoning Ag**e**nts for **L**ongitudinal care），这是一种用于心理健康随访护理中纵向预后推理的多智能体推理支架。我们与12名精神科医生开展了形成性研究，以识别精神科随访所涉及的关键推理功能，并将这些功能转化为系统设计。CHORAL构建纵向轨迹，分别对抑郁、焦虑、压力和睡眠进行预后评估，识别信息缺口，并通过跨领域综合整合所得的评估结果。在接触其他领域评估之前，系统会先完成每个领域的评估，从而使解释差异能够在综合之前显现。随后进行的审查和修订会使支持性证据、分歧及变化保持明确且可审查。

我们探讨以下研究问题：

- **研究问题1.** CHORAL 的架构组件如何提升纵向患者生成健康数据的预后性能、证据使用和引用有效性?

- **研究问题2**。 临床医生如何看待 CHORAL 在解读纵向患者生成健康数据并制定随访计划方面的有用性?

- **研究问题3**。 CHORAL 如何支持临床医生质疑预后判断，并决定哪些内容可以依赖、哪些内容需要核实?

为回答这些问题，我们开展了两项互补性评估。首先，我们在一个纵向心理健康数据集上，将 CHORAL 与多数类基线、持久性基线和单智能体基线进行比较，并开展组件消融实验（研究问题1）。随后，我们对 22 名精神科临床医生开展了受试者内研究，将 CHORAL 与单智能体基线进行比较，以考察临床有用性、随访支持、推理审查和依赖支持（研究问题2–研究问题3）。

我们的研究结果表明，CHORAL 的平衡准确率与持久性基线相当（.673 对 .672），同时检测出了持久性基线未能检测到的 40% 的从低到高的恶化病例。与单智能体基线相比，CHORAL 还取得了更高的平衡准确率（.673 对 .639）和更广泛的证据渠道覆盖率（89.2% 对 70.1%）。在临床医生研究中，CHORAL 尤其受到认可，因为它支持证据评估、呈现替代性解释，并为质疑系统判断提供具体依据。它还帮助临床医生区分系统输出中哪些方面可以依赖，哪些方面需要进一步核实。在 CHORAL 条件下，跨领域解读的评价更高，但经过Benjamini–Hochberg校正后，这一差异不再具有显著性。相比之下，这些益处并未明确延伸至随访优先级排序或工作流集成。基于这些发现，我们讨论了以临床医生为中心的人工智能系统的设计启示，以支持心理健康随访护理中可检查且可质疑的纵向预后推理。

总之，我们的贡献如下:

- 通过对 12 位精神科医生开展形成性研究，我们推导出纵向、多领域预后推理的设计要求，并将其落实到 CHORAL 中。CHORAL 保留各领域独立的评估，区分可检索的信息缺口与无法通过外部途径解决的信息缺口，并使综合与修订过程可追溯，以供临床医生审查。

- 通过系统级评估和临床医生评估，我们表明，CHORAL 能够支持更广泛的证据使用，并帮助临床医生识别替代性解释、找到提出异议的依据，以及区分哪些内容需要进一步核实。

- 我们提出了设计启示：如何呈现具有实质影响的审议，同时将推理深度与呈现深度区分开来。

2 背景与相关工作

本节回顾有关纵向预后推理、人工智能临床推理支持，以及用于构建临床判断的多智能体系统的现有文献。

2.1 心理健康随访中的预后推理

预后推理关注患者病情可能如何演变，以及这一预期病程应如何指导后续护理 [23, 49]。在随访护理中，这类判断具有纵向性且是暂定的：临床医生会比较

当前状态与患者既往病程，评估变化的方向与持久性，并在新证据出现时修正先前的解释 [22, 28]。

纵向患者生成的健康数据（PGHD）能够捕捉临床就诊之间的行为变化，从而为这一过程提供更多证据。先前的传感研究已发现，活动能力及其他行为信号与抑郁症状相关 [9, 54], ；近期的数字表型分析研究则显示，个体之间以及同一个体随时间推移，其行为和生理模式均存在显著差异 [16, 39, 76]。综合而言，这些研究表明，PGHD能够为了解患者行为状态随时间的变化提供有用证据。

然而，在心理健康领域，解读这些行为变化的临床意义仍然十分困难。精神症状经常相互重叠并同时出现，因此很难仅从单一症状的角度解读观察到的变化 [8, 29]。同一种行为表现可能支持多种与症状相关或非临床的解释。评估此类变化的临床意义，还需要考虑其与日常功能的相关性 [4, 66]。主观体验和情境背景对于区分各种可能的解释也很重要，但被动感测或结构化数据未必能捕捉到这些信息。

因此，使用患者生成健康数据进行预后判断，需要结合患者既往病程来解释变化，考虑多种合理的解释，识别缺失的情境信息，并随着新证据的出现不断更新判断。在时间受限的医疗护理中，跨异质性纵向数据执行这些步骤，可能会带来 substantial 的解释负担 [59, 63]。这就产生了对相关工具的需求：这些工具应支持上述中间推理步骤，而不仅仅是呈现纵向数据或最终预测结果。

2.2 用于预后推理的人工智能支架

以往的人工智能和人机交互系统通过将多模态传感流整理为模式、叙事和可视化，帮助用户解释纵向健康数据 [13, 75, 83]。这些系统使行为轨迹更易于理解，但主要侧重于理解已观察到的或当前的状态。

相关研究利用纵向感知数据和健康记录数据，估计未来的心理健康状态和临床状态 [5, 36, 60, 82]。这些方法将纵向分析从理解已观察到的状态扩展到预测未来结局。然而，即使预测结果附有支持性证据，临床医生仍需判断应如何解释这些证据、哪些替代性解释仍然合理，以及在形成临床判断之前可能还需要哪些额外信息。

近期的人机交互研究日益关注人工智能如何支持人类推理过程，而不仅仅是提供最终建议。Sivaraman等人将人工智能提供的信息概念化为推理线索，而Zhang等人则对比了以建议为中心的人工智能与在整个决策过程中提供持续支持的人工智能 [59, 80]。鉴于有证据表明，仅仅让医生获得功能强大的大语言模型，并不一定能比传统资源更有效地提升其诊断推理能力，这一区分尤为重要 [20]。总体而言，这些发现促使人们开发能够支持判断形成过程的人工智能，而不是简单地给出答案。

临床推理支持还引发了一个问题：临床医生能否检查并质疑人工智能支持的判断所依据的基础。可解释人工智能研究强调了让算法决策变得易于理解的重要性 [44], ，而可质疑人工智能则主张，用户应能够批判性地参与算法判断并对其提出质疑 [2, 41]。近期的精神病学人工智能系统 Heart2Mind 通过将解释与临床医生与人工智能对话相结合，来支持对模型预测的检视与质疑，从而将这一方向付诸实践 [48]。然而，这项工作主要聚焦于审查人工智能输出及其依据，对判断如何通过基于纵向证据的中间推理逐步形成关注较少。

2.3 用于临床推理的多智能体人工智能

近期临床人工智能研究越来越多地采用多智能体系统，通过专业化角色和不同视角来组织复杂推理 [11, 35, 62]。早期方法主要强调专业化、任务分解，以及跨专家角色的汇总 [35, 62, 69, 70]。后续系统则通过自适应配置、主智能体编排或元层级协调，明确构建协作与综合结果 [11, 35, 40, 71]。这些研究表明，研究重点已从分配专业化角色转向协调如何整合分布式判断。

在心理健康领域，多智能体系统主要对不同的临床职能或阶段进行分工，包括访谈、筛查、诊断评估和患者沟通 [7, 21, 32, 61, 73]。相比之下，多领域预后要求在综合之前，从多个症状领域视角对同一纵向轨迹进行评估，因为同一行为模式对于抑郁、焦虑、压力和睡眠可能具有不同含义 [8, 12, 29]。因此，预后推理需要在综合之前保留领域特定评估，同时允许随着新证据的出现重新审视暂定判断 [22, 28]。

因此，挑战不仅在于生成多个评估，还在于构建一种机制，使这些评估以证据为依据、跨领域综合，并通过审查重新考量。既有系统已支持对被动传感数据进行多大语言模型意义建构（例如 GLOSS [13]）、对多模态心理健康数据进行面向临床医生的叙事解读（例如 MIND [83]），以及围绕临床记录开展多智能体协作（例如 ColaCare [71]）。相比之下，CHORAL 专注于保留预后解读如何基于同一纵向证据不断演变：在综合之前保留各个独立的领域评估，根据能否从记录中解决来区分信息缺口，并确保后续审查或修订仍可追溯至临床医生。

3 系统设计

CHORAL 的设计依据是形成性研究所发现的临床推理需求和系统支持需求。本节首先介绍形成性研究及由此得出的设计要求，随后介绍为实现这些要求而开发的 CHORAL 架构。

3.1 形成性研究

我们的目标是设计一个系统，支持在心理健康随访护理中利用患者纵向信息进行审慎的临床推理。因此，我们开展了一项形成性研究，以了解临床医生如何解读纵向变化并将其转化为临床判断，以及系统如何支持这一过程。

3.1.1 研究设计。 我们对从韩国各地大学医院和诊所招募的12位精神科医生进行了半结构式访谈（4名女性；年龄 M=44.5, SD=6.43）。临床经验定义为获得专科认证后的执业年限：两位参与者有0–5年经验，两位有6–10年经验，另有八位超过10年。参与者通过滚雪球抽样招募，先从一小群精神科医生开始，再通过专业人士推荐扩大样本；这种方法常用于在人机交互研究中招募难以接触的临床人群 [31]。

每次访谈约持续 60 分钟，均在线进行。访谈主要围绕两个主题展开：（1）临床医生在心理健康随访护理中的推理方式；（2）系统如何支持基于患者纵向信息进行推理。经参与者同意，访谈均录音，每位参与者获得约 150 美元的补偿。本研究经机构审查委员会（IRB）批准，所有参与者均在参与前提供了书面知情同意。

表1。形成性发现、代表性参与者引语及相应的设计要求。

形成性发现	代表性引述	设计要求
临床医生考察了变化 相较于患者既往病程轨迹 轨迹	**P3:** “我首先会问，这是否确实与 该患者平时的情况不同……” **P8:** “我首先会看变化方向， 看患者是在恶化、好转，还是 只是短暂波动。”	**DR1.** 先建立纵向 轨迹，再进行 领域特定评估
同一种行为变化 会在多个 领域	**P4:** “大多数患者并不能简单地归入单一 诊断……” **P9:** “当我看到活动减少时，我不会 立即想到抑郁……我通常会查看 多个症状领域，然后再决定 变化的含义。”	**DR2.** 评估纵向 跨多个 领域
含糊不清的解释 促使临床医生寻求 缺失信息	**P6:** “如果有些事情对不上，我可能会查看 既往病程，或者询问家庭成员……” **P10:** “有时数据会告诉我有些情况发生了变化，但不会说明原因……我会想核实 在做出判断之前， 那段时间发生了什么 判断。”	**DR3.** 识别信息 缺口并支持针对性 随访
临床医生整合了 领域特定评估 在比较它们如何 相互补充或彼此冲突 时	**P1:** “我把这些片段拼凑起来，以理解 整体情况……” **P8:** “我会观察它们如何相互衔接，以及它们在哪里 冲突，以及哪些顾虑应当优先处理。” **P7：** “我可能会认为抑郁有所好转 而睡眠问题却在恶化……单一 领域并不能让我了解全貌。”	**DR4.** 整合不同视角 同时保留可追溯性
就诊时间有限 优先呈现信息， 详细内容可按需查看	**P1:** “根本没有时间通读 冗长的文本。” **P11:** “如果摘要能……效率会更高 首先出现……” **P12：** “如果系统一次性给我所有信息，我 可能不会在随访就诊期间阅读…… 我更希望先看到主要问题……”	**DR5.**为信息排序 以适应时间受限的 随访

两名研究人员使用主题分析法，分别对访谈记录进行编码。随后，研究人员通过亲和图法和讨论，对逐渐形成的主题进行整理和完善，直至达成共识。通过这一分析，我们识别出纵向随访护理中的临床推理实践，以及临床医生对系统支持的需求。

3.1.2 研究发现与设计要求。 我们的分析确定了纵向随访护理中的四项关键推理功能：(1) 追踪纵向变化，(2) 评估多个领域的纵向变化，(3) 识别信息缺口，以及 (4) 整合领域特定评估以支持随访。临床医生还强调，系统应支持而非取代临床判断，使系统生成的判断依据可供审查，并以便于在有限随访时间内审查的形式呈现信息。基于这些发现，我们归纳出五项设计要求，并在表1中通过代表性参与者引语加以概述。

DR1：在进行领域特定评估之前，先建立纵向轨迹。临床医生首先会根据患者以往的状态评估变化，重点关注变化的方向和持久性基线，然后再进行领域特定解读。有些人将这一顺序与 SOAP式记录联系起来；在这种记录方式中，观察结果会先于临床评估进行整理 [72]。因此，系统应先建立患者层面的纵向轨迹，再进行领域特定评估。

DR2：评估多个领域的纵向变化。临床医生从多个领域特定视角审视同一行为变化，而不是直接将其归入单一解释。因此，系统应在病例层面进行整合之前，保留各自独立的领域特定评估。

DR3：识别信息缺口并支持针对性随访。当观察到的变化难以解读时，临床医生会从既往病史、患者或其他来源寻求更多情境信息。因此，系统应在可能的情况下检索相关纵向证据，并呈现尚未解决的信息需求，以便后续临床随访。

DR4：整合领域特定评估，同时保留可追溯性。临床医生通过比较各领域特定评估如何相互补充或冲突，同时保留其支持性证据，形成总体理解。因此，系统应整合这些评估，同时保留每种解释所依据的证据和推理。

DR5：在时间受限的随访中优先呈现信息。临床医生强调，在有限的随访时间内，需要优先关注近期变化、主要问题和未解决事项。因此，系统应首先呈现关键信息，同时确保用户可按需查看支持性证据和推理。

这些设计要求为下一节所述 CHORAL 的推理机制和界面设计提供了依据。

3.2 CHORAL系统

基于形成性研究中归纳出的五项设计要求，我们开发了 CHORAL，这是一个用于支持基于纵向患者生成健康数据开展临床推理的多智能体系统。图1概述了系统架构和推理流程。CHORAL 并不生成单一的端到端判断，而是将推理过程组织为多个阶段。患者代理首先根据患者生成健康数据构建共享的纵向轨迹。随后，领域智能体分别从抑郁、焦虑、压力和睡眠的角度，独立分析同一纵向轨迹。当需要补充信息时，CHORAL 会从现有纵向病史中检索相关信息，或保留尚未解决的信息需求，以供后续临床随访使用。编排器在病例层面整合领域特定的预后评估，同时保留这些评估所依据的证据以及不同视角之间的差异。最后，临床仪表板首先呈现主要发现，同时保留详细证据和推理过程，以便进一步查阅。实现细节和提示模板见附录C。

3.2.1 患者代理。为支持 DR1，患者代理根据纵向患者生成健康数据构建共享的纵向轨迹。在形成性研究中，临床医生强调，应考察患者的行为相较于其通常状态发生了怎样的变化、这种变化是暂时的还是持续的，以及它是否可能与日常功能相关。患者代理不会直接将原始PGHD与特定领域联系起来进行解读，而是先将行为变化及其与功能相关的背景整理为患者层面证据，供后续从多个领域特定视角进行解读。

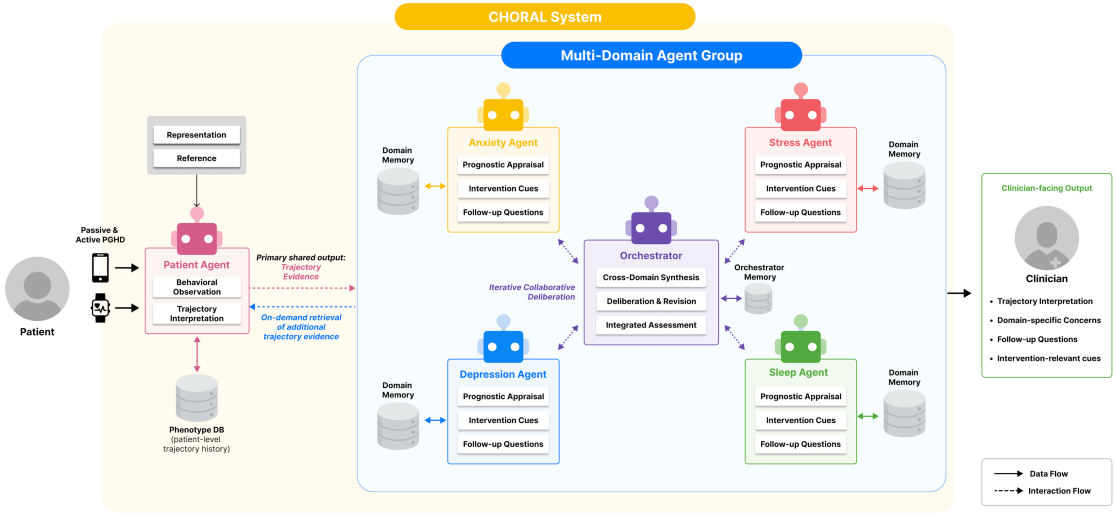


图1。CHORAL系统架构和推理流程概览。

构建患者特定的纵向轨迹，需要考虑发生了哪些变化、变化方式如何、当前状态与患者通常模式有多大差异，以及当前变化与患者既往病程有何关联。我们围绕三种机制组织这一过程：表征、参照和情境。

表征：显式行为轨迹表征。纵向患者生成健康数据由活动、活动能力、睡眠和手机使用等异质信号构成。仅凭原始值无法一致地描述患者的行为状态或其随时间的变化。既有关于感知表征的研究同样表明，除了原始传感器值之外，表示时间模式和变化也十分重要 [78]。因此，我们根据统计特征、随时间的变化以及日内模式来组织患者生成健康数据。

对于每个患者生成健康数据流，统计特征使用均值、标准差、最小值和最大值、中位数以及首次和最后一次观测值，总结观察期内的整体状态和变异性。结构特征使用斜率、趋势方向、变化量、百分比变化以及日间变异性，描述变化的方向和幅度。时间特征依据既有感知研究，使用四个6小时的时段捕捉日内行为模式：夜间（00:00–06:00）、上午（06:00–12:00）、下午（12:00–18:00）和晚间（18:00–24:00） [74]。

为呈现与功能相关的背景，我们还依据世界卫生组织残疾评定量表（WHODAS） [66]，通过在现有感知特征与功能相关领域之间建立概念映射，对这些行为特征进行组织。完整的特征到领域映射见附录B。该映射有助于解读可能与功能相关的行为变化，而非推断 WHODAS 评分或判定功能损害。

参照：相对于基线的轨迹构建。同一行为值在不同个体身上可能具有不同含义，因为每位患者都有不同的通常模式。因此，当前行为需要结合个体内变化来解读，而不能仅依据绝对水平 [18]。形成性研究中的临床医生也表示，在评估是否发生了有意义的变化时，会将当前状态与近期行为以及更稳定的历史模式进行比较。

对于每一轮推理，当前状态均在7天窗口（$t-6$ 至 t）内进行汇总。对于被动感知和EMA数据，近期基线根据当前窗口之前的28天期间计算得出，并排除当前窗口（$t-34$至 $t-7$）。对于被动感知特征，稳定基线根据当前窗口之前所有可用的感知历史计算得出，同时排除与满足基线排除标准的EMA评分相关的感知时段（压力 ≥ 2、PHQ-2 ≥ 2或GAD-2 ≥ 2）。对于EMA测量，稳定基线由当前窗口之前所有可用的EMA响应构成，不进行基于症状的筛选。这些参照时段反映了个体内的短期和长期变化。在可获得的情况下，同时使用近期基线和稳定基线；如果其中任一基线无法构建，则省略该比较，而不进行插补。

情境：患者特异性历史锚定。在呈现当前行为状态及其相对于基线的变化后，患者代理会比较多个时间窗口中的行为模式，以确定当前轨迹的方向和情境。CHORAL 将这些纵向记录累积在表型数据库中，每条记录都包含自报告症状状态、显著行为模式、相对于基线的变化以及患者层面总结。每个证据项都会被分配一个唯一标识符，注明患者、证据渠道、周次和具体项目（例如 `[UID179-SLEEP-23#03]`），使智能体能够在推理时直接引用支持性证据。之后的推理过程中，还可以重新查阅这些累积的记录，作为历史背景。

3.2.2 领域智能体。为应对 DR2，CHORAL 使用不同的领域智能体，从各自不同的领域特定视角评估同一条共享轨迹。在形成性研究中，临床医生指出，同一种行为变化可能支持多种解释，这表明在病例层面整合之前，有必要保留各自独立的评估。这促使系统在综合之前先进行领域特定评估。

我们设置了四个领域智能体，分别对应抑郁、焦虑、压力和睡眠。这些领域之所以入选，是因为它们代表了心理健康与福祉中重要且密切相关的方面 [14, 43]，也是 CHORAL 实现多领域推理所依托的领域。

独立专业化：领域特定预后评估。每个领域智能体都会从自身领域的角度，独立评估共享轨迹。该智能体会考虑当前行为模式可能与该领域有何关联、是否引发领域特定的预后关注点、有哪些证据支持或削弱该评估，以及可能相关的随访或管理考量。关于领域特定的背景知识，抑郁、焦虑和睡眠智能体参考基于 DSM 的标准 [4],，而压力智能体则依据成熟的压力框架 [43]。这种结构使同一行为证据能够从多个角度分别解读，同时在整合之前保留各领域特定的评估。该设计借鉴了以往的医疗多代理系统：在这些系统中，具备不同临床专业知识的专科代理会先独立审查同一病例，再整合各自的评估 [35, 62, 71]。

领域记忆：纵向领域特定推理。领域特定预后评估并非一次性判断，而是通过将较早的评估与随后观察到的症状状态联系起来进行更新。临床推理同样将初始解读视为一种临时性判断，随着新证据的出现，该判断可能会被审查或修订 [22, 28, 49]。在每一轮推理中提供完整的纵向病史，可能会使既往信息难以得到一致使用。以往的大语言模型智能体研究通过将过去的信息存储在显式记忆中，并在后续推理期间检索选定记录来解决这一问题 [79, 81]。

在 CHORAL 中，每个领域代理都会显式记忆此前的领域特定推理。每条记忆记录都保存患者的病程轨迹以及该代理当时的领域特定评估，之后还会关联随后观察到的症状状态，并对先前的评估进行反思。选取的过往记忆可

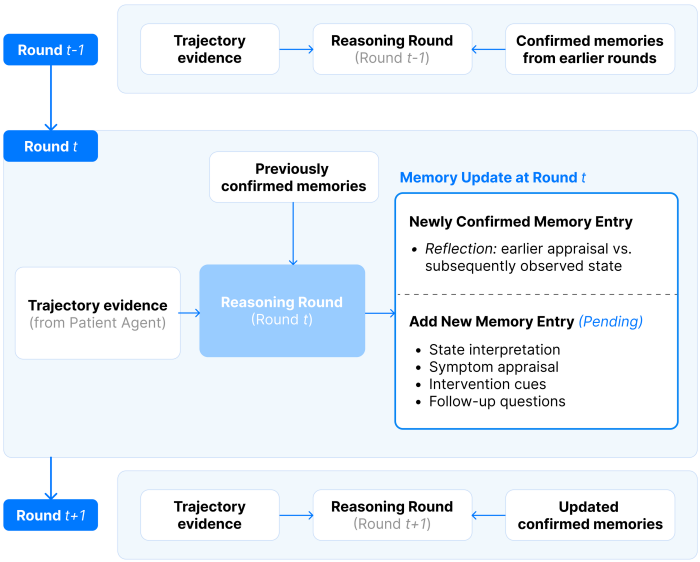


图2. 跨推理轮次的领域智能体记忆。当前轨迹证据和近期已确认的记忆为每次新的预后评估提供信息，随后再根据观察到的结果和反思进行更新。

在后续轮次中检索，使当前评估能够结合患者既往病程进行考量。领域记忆包含两个过程：记忆构建，用于保存每个推理轮次的结果，并在随后获得症状状态后更新先前记录；以及记忆检索，用于将选取的过往记录引入当前推理轮次。

记忆构建。 如图2所示，每个推理轮次都会更新先前待处理且其目标症状状态现已可用的记忆条目，同时为当前轮次创建新的条目。领域代理首先从自身领域视角解读共享轨迹，并将结果存储为新的记忆条目。每个条目包含四个组成部分：状态解读、领域特定预后评估、干预提示和随访问题。状态解读概括对当前行为模式和自我报告的症状状态的理解，而领域特定评估则概括该领域的预后关注点及其支持性证据。干预提示总结相关的管理考量，随访问题则指出之后需要澄清的信息。

由于创建条目时尚无法获得目标症状状态，新条目最初会被存储为待处理。当预定义的预测时限过后，与较早评估相对应的目标症状状态可用时，系统会将较早的评估与观察到的状态进行比较，并添加一条反思。例如，在预测时限为两周的情况下，第 t 轮生成的评估会一直处于待处理状态，直到第$t+2$ 轮的症状状态可用。反思会记录较早的判断在哪些方面与后续病程一致或存在差异，哪些证据在解释该病程时仍然有用，以及哪些解读部分的支持力度较弱。这种设计借鉴了基于反思的智能体方法：审查较早的推理结果，并在后续推理中重新利用该审查结果 [57]。

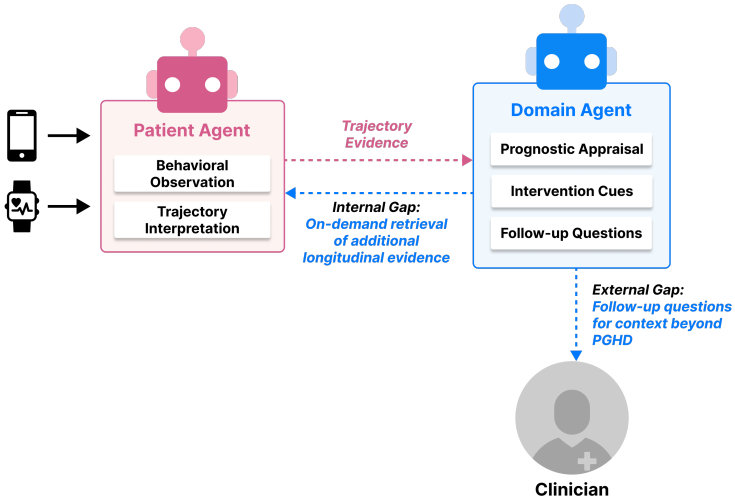


图 3. CHORAL 信息缺口处理流程概览。

添加反思后，相应条目会成为一条已确认 记忆，而较新的条目由于其目标症状状态尚未被观察到，仍保持待处理状态。随着轮次推进，领域记忆会积累一段相互关联的历史，其中包括患者背景、领域特定评估、后续症状状态和反思。

记忆检索。在每个推理轮次中，领域代理会从较早轮次中检索已确认记忆，其中包含其先前的评估、随后观察到的症状状态以及相应的反思。这使代理能够考虑其先前的判断与患者后续病程之间的关系。

CHORAL 检索近期的已确认记忆，而不是全部记忆历史。以往的大语言模型代理工作同样将近因性作为重要的检索信号 [51, 81]。在我们的实现中，每个领域代理都会在当前推理轮次接收最近的三条已确认记忆。

3.2.3 信息缺口处理。为落实 DR3，CHORAL 会识别澄清领域代理初始评估所需的额外信息，并确定这些信息能否从纵向患者记录中检索，还是应留待临床随访（图 3）。我们将前者称为内部缺口，即通过从现有纵向记录中进行额外检索即可解决的缺口；将后者称为外部信息缺口，即需要超出可用患者生成健康数据（PGHD）的情境信息，因此保留至临床随访中处理。在形成性研究中，临床医生描述了这样一种做法：当观察到的变化的含义不明确，或与患者的解释不太相符时，先识别缺失的信息，然后在随访期间重新了解患者既往病程或寻求更多情境信息。临床推理文献同样将初始解读视为暂定结论，并认为随着更多相关信息的获得，该解读应进一步完善 [22, 33]。

识别信息需求。 在生成初始评估后，每个领域代理都会审查其解读中哪些部分仍不明确，以及哪些额外的患者层面信息有助于澄清这些部分。该代理会结合其解读以及支持性证据和削弱性证据，识别当前判断背后的不确定性，并将减少该不确定性所需的信息明确为一项信息需求。

这些信息需求分为两种类型。内部缺口 可以通过现有的纵向病史来解决，例如当前的模式是否曾经出现过，或类似的行为变化过去是如何伴随症状状态出现的

。外部信息缺口 无法通过当前的患者数据来解决，例如变化的情境原因或患者的主观解释，因此会留待后续临床随访期间直接评估。这体现了临床问题解决中的选择性信息收集：临床医生会寻求与当前解读相关的信息，而不是再次审查所有可用信息 [33]。

基于查询的检索与随访。 对于内部缺口，相应的信息需求会转化为系统级查询 并发送给患者代理。患者代理不会再次总结全部病史，而是从纵向记录中有选择地检索与查询相关的患者层面证据，并将其返回给领域代理。这遵循了以往从更大规模患者记录中提取临床相关信息的基于检索的方法 [3]。例如，它可以核查当前行为模式是否曾经出现过，或类似模式过去如何与症状状态共同出现。

对于外部信息缺口，相应的信息需求会保留为临床医生层面的随访 ，而不会在系统内得到解决。这可能包括变化发生时的情境背景、患者的主观体验，或传感数据或自我报告数据未能捕捉到的其他信息。这些未解决的信息需求会以随访问题的形式呈现，临床医生可在后续就诊时与患者讨论。

患者代理返回的证据会传回领域代理，以便进一步完善初始评估。如果证据支持现有解读，则可保留或强化该评估；如果证据指向不同方向，则可予以修订；如果问题仍未解决，则可将其标记为不确定。这样便能将当前评估与纵向患者记录中已有的信息重新关联起来，并且整个过程仅限于单次查询—响应—完善循环。

3.2.4 编排器。 为回应 DR4，CHORAL 使用编排器，将抑郁、焦虑、压力和睡眠领域智能体的评估结果整合起来，形成对患者整体状态的综合评估。在形成性研究中，临床医生表示，他们会比较不同领域的证据和解释来形成整体判断，而不是依赖单一解释。以往的多智能体医疗系统也采用过协调智能体，综合由具有不同临床视角的专科代理独立生成的判断 [35, 62, 71]。

跨领域综合。 编排器会检查所有四个领域智能体给出的领域特定预后评估及支持性证据，以了解不同领域的问题与证据之间的关联。它会识别得到多个领域共同支持的行为模式或预后问题，同时保留特定于某个领域或在不同领域之间存在差异的解释。基于这些评估结果，编排器会生成一份初步综合评估，明确保留每种解释由哪个领域提出以及其依据的证据，从而保留各领域的推理轨迹。

领域审查与可追溯修订。 编排器生成初步综合后，每个领域代理都会审查其领域特定评估及支持性证据是否在综合结果中得到了恰当体现。代理会检查自己的解释是否被过度简化，或是否以与支持性证据不一致的方式纳入综合结果。如需修订，代理会提供反馈，说明哪些内容应重新考虑以及原因。例如，代理可能会指出，某一领域的顾虑被弱化得过头，或不同领域之间的不同解释被合并，导致重要差异消失。

编排器根据这些反馈修订初步综合结果，随后领域智能体会再次审查修订后的综合结果。参照以往采用有限轮次多智能体审议、逐步完善集体评估的工作，该审查与修订流程最多重复三轮。 [35],

审查流程完成后，编排器将修订后的综合结果定稿为综合评估。最终结果包括各领域的评估、主要支持性证据、各领域之间的一致意见与差异，以及剩余不确定性和后续跟进需求。领域代理的原始评估和审查反馈也会予以保留，以便临床医生追溯形成最终评估的证据和审查流程。

编排器记忆：纵向综合推理。编排器会记忆在早期推理轮次中如何对多个领域进行综合解读。在完成领域审查和必要的修订后，每个推理轮次的最终综合评估都会作为一个记忆条目存储。随后，当新的症状状态可用时，系统会更新先前的条目，记录早期最终综合评估与患者后续症状变化过程之间的关系，并将其标记为已确认。这既保留了早期的综合评估，也保留了相应的跨领域解读与后续症状变化过程之间的关系。

在新的推理轮次中，编排器会检索近期的已确认记忆，并结合先前的最终综合评估以及随后的症状状态进行考量。这使其不仅能够考虑当前领域代理的输出，还能考虑同一患者此前出现的跨领域模式，以及这些模式之后如何发展。在我们的实现中，会检索最近的三个已确认记忆，以便将当前的综合评估与患者近期的推理历史联系起来。

3.2.5 临床仪表板。 为满足 DR5，CHORAL 通过临床仪表板呈现前述推理结果，帮助临床医生快速识别关键发现，并将其用于后续护理。在形成性研究中，临床医生指出，就诊时间有限，难以从头开始审查所有纵向数据和完整的推理轨迹。因此，CHORAL 优先呈现患者当前状态和近期变化、支持评估的主要证据，以及与随访相关的信息，而不是一次性展示完整的推理过程（图4）。如需进一步检查，临床医生可以查看各领域代理的评估、编排器的审查过程，以及底层推理轨迹（图5）。

4 评估方法

我们从两个方面评估了 CHORAL。首先，我们开展了定量系统评估，以考察其设计组件如何促进基于纵向患者生成健康数据的预后性能和证据依据。其次，我们开展了临床医生比较研究，以考察 CHORAL 的多智能体编排如何支持预后推理、随访规划、推理审查和依赖支持。

4.1 定量系统评估

为对 CHORAL 进行定量评估，我们将完整系统与参考基线进行了比较，并开展了消融研究，以考察其架构组件如何影响预后性能、证据使用和引用有效性。

4.1.1 数据集与评估协议。 我们使用大学生经历研究（CES）数据集 [47],对 CHORAL 进行了评估。该数据集包含一个纵向大学生队列的被动传感数据，以及重复测量的自我报告心理健康测量。参与者每周完成生态瞬时评估（EMA），使用 PHQ-2 测量抑郁症状，使用 GAD-2 测量焦虑症状，并评估感知压力。这些重复的自我报告用于定义定量评估的参照症状状态。

为提供充分的纵向背景，我们纳入了最早符合条件的连续16周时间段内的参与者；该时间段内所有计划进行的自我报告均已提供，且每周均满足传感数据质量标准。根据 CES 每日质量指标 [34], ，当某传感器数据流至少有 60% 的 24

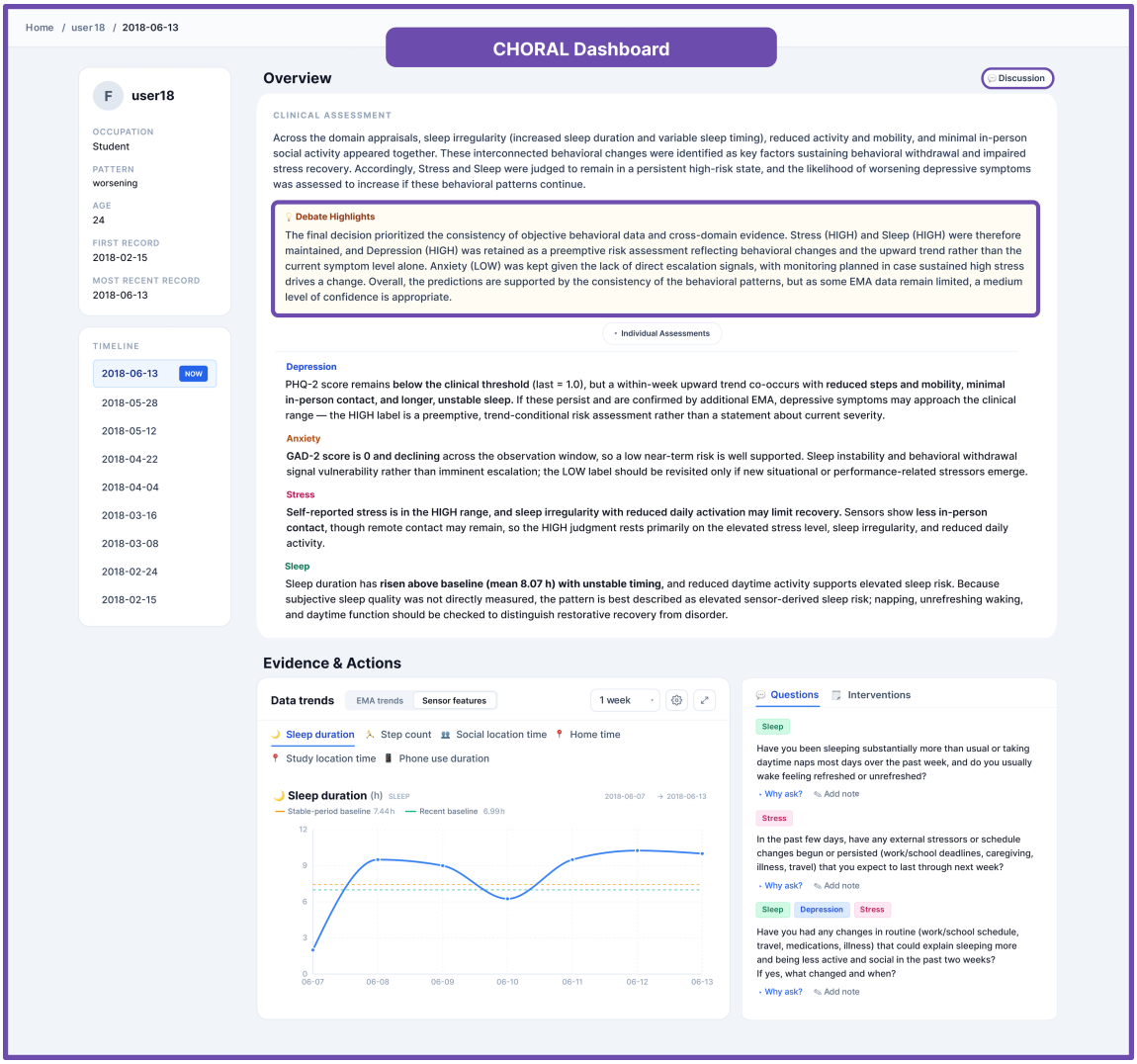


图4. CHORAL临床仪表板概览。该界面优先呈现关键发现、支持性证据和后续信息，同时提供对领域代理评估及编排器审查流程的访问。

个小时区间包含数据时（鉴于质量指标为整数，相当于至少 15 小时），则视为该日完整；每周的活动数据和位置数据分别评估，且各自至少需要五个完整日。筛选后，共有 74 名参与者纳入分析。

对于每个目标评估，我们为抑郁、焦虑、压力和睡眠定义了二元的高状态和低状态。抑郁和焦虑分别使用 PHQ-2 $\geq$ 3 和 GAD-2 $\geq$ 3进行二值化 [37, 38]。对于压力，我们将五点评分量表中得分为 $\geq$ 4 的情况定义为高状态，这对应于高于中性中点的回答，表明存在感知压力。由于 CES 不包含每周自我报告睡眠测量，我们根据睡眠时长、时间安排和不规律性，通过被动感知推导睡眠结果；因此，睡眠结果应解读为对未来传感器推导的睡眠状态的预测，

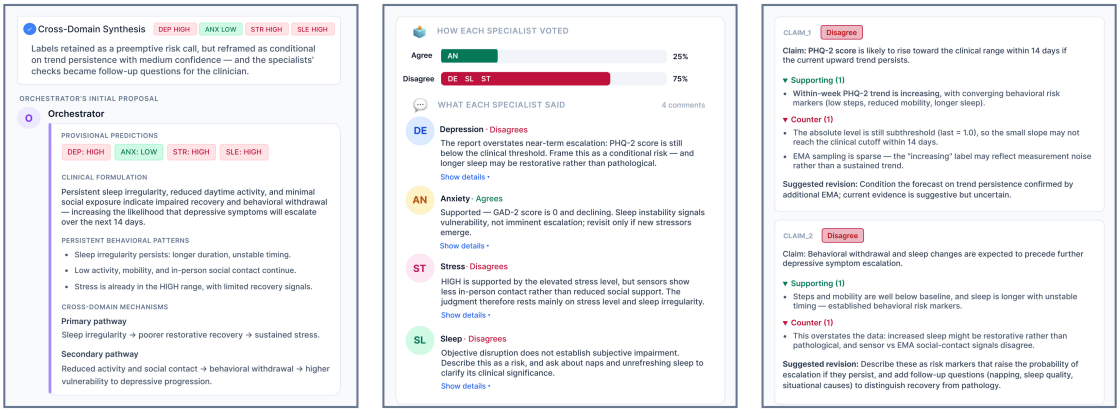


图 5. CHORAL 审议界面的详细视图：(A) 编排器最初生成的跨领域综合结果；(B) 领域代理的审查结果，显示其与综合结果一致或存在分歧；(C) 按主张展示的审查详情，包括支持性证据、反面证据和建议的修订。

而非对自我报告的症状状态的预测。睡眠标签构建的详细信息见附录A。

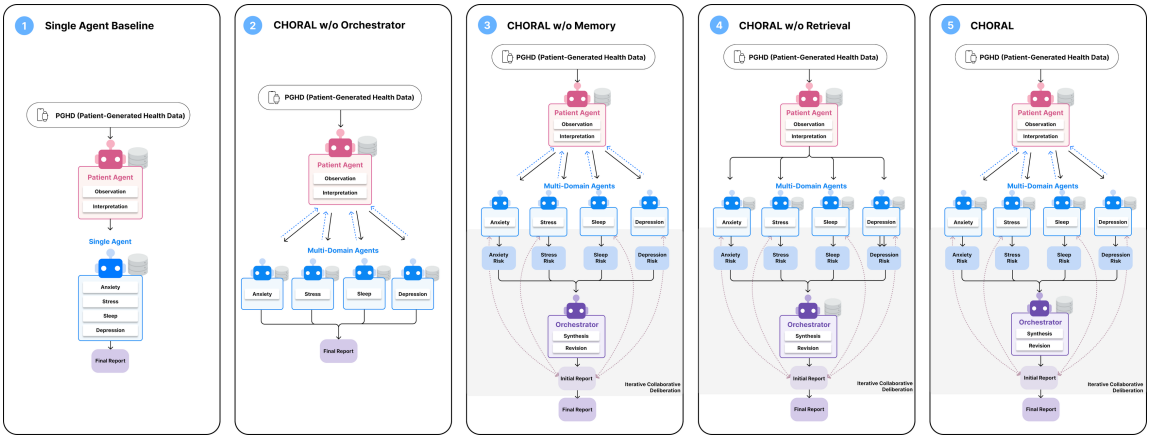
在评估中，前四周用于建立患者特异性基线背景。从第 5 周开始，系统利用可用的纵向患者生成健康数据，估计各领域约两周后的状态。我们选择这一时间跨度，以近似门诊随访就诊之间数周的间隔；例如，基于测量的抑郁症护理在急性期治疗期间通常每 2–3 周进行一次随访就诊 [25]。在每个推理轮次中，系统只能访问截至该轮次可用的信息；未来目标状态在可用于后续结果确认之前不会向系统披露。在这一两周的时间跨度下，每项评估都会保持待定，直至其目标状态变得可用。

主要分析包括第 7 周至第 16 周的目标评估。例如，在第 5 周生成的评估，是根据大约第 7 周的症状状态进行评估的。在 74 名参与者中，这产生了 740 个病例周和 2,960 个领域级评估案例。

4.1.2 基线与消融。 我们将完整的 CHORAL 与三个基线进行了比较。多数类基线 始终为每个领域分配多数类。持久性基线 假设患者未来的目标状态将与当前观察到的状态保持一致。

此外，我们构建了一个单智能体基线，作为更强的架构对比；该基线保留了共享的患者代理轨迹、领域知识、目标定义、输出要求以及可比的纵向记忆。在接收与 CHORAL 所使用的相同患者代理轨迹后，单个大语言模型对抑郁、焦虑、压力和睡眠进行评估，并在一次生成中整合四个领域特定评估。因此，在没有独立领域智能体、按需检索、评估修订，或编排器主导的综合、审查与修订的情况下，它执行了相同的总体预后任务。

图 6 总结了经评估的系统配置。为考察 CHORAL 各个组件的贡献，我们评估了三种消融变体：



Fi图6. 用于定量评估的系统配置 估。

- **无按需检索**：移除了领域代理从患者纵向记录中请求额外证据的机制
- **无记忆**：移除领域记忆和编排器记忆，同时保留所有条件共用的患者代理轨迹和纵向患者证据
- **无编排器**：保留领域特定的预后评估，但移除由编排器执行的跨领域综合、领域审查和条件修订

4.1.3 评估指标。 我们从两个维度评估CHORAL：预后性能和证据依据。对于预后性能，平衡准确率和恶化检测是互为补充的主要指标。平衡准确率考虑了高状态和低状态之间的类别不平衡，而恶化检测则反映了预测低状态到高状态转变的能力 [15]。精确率、召回率和F1值作为补充指标进行报告，其中高状态为正类。平衡准确率、精确率、召回率和F1值在四个领域之间进行宏平均，而恶化检测则汇总所有观察到的低到高转变。对于当前处于低状态的案例，警报 指预测未来将处于高状态；警报精确率 指警报后发生低到高转变的比例，而误报率 指被错误预测为高状态的低到低案例比例。对于关键比较，在参与者层面通过2,000次自助法重采样估计95%置信区间，并保留每位抽样参与者中的所有病例周和领域预测。

对于证据依据，我们评估了每种条件下的最终输出：编排条件下的综合评估、无 Orchestrator 条件下各最终领域智能体评估的拼接结果，以及单智能体基线下的最终评估。证据渠道覆盖率 是指每个病例周至少被引用一次的可用（非空）渠道所占的比例，这些渠道取自七种渠道类型（ACT、EMA、LOC、PHONE、SIG、SLEEP 和 SOCIAL）；每个渠道最多计数一次。无效证据引用率 是指引用出现次数中与相应患者记录中的证据标识符不完全匹配的比例，重复标签分别计数。计数跨病例汇总，无效引用经过人工检查。这些指标反映的是参照广度和标识符有效性，而非临床相关性或语义正确性。

4.2 临床医生评估

我们与临床医生开展了一项用户研究，以评估 CHORAL 相较于单智能体方法在预后推理、随访规划、推理审查和依赖支持方面的表现。

4.2.1 研究设计。 我们采用被试内设计，比较单智能体条件与 CHORAL。单智能体接收相同的患者代理轨迹，并在一次生成中给出四项领域评估和一项综合评估。除独立的领域智能体外，CHORAL 还包含按需证据检索与评估修订、由编排器主导的综合、审查和修订，以及单智能体条件中不具备的中间推理视图。因此，临床医生研究比较的是两种系统配置的整体，而非单独考察智能体组织方式的影响。

我们从 CES评估集 中选取了四个病例，以代表不同的纵向病程：恶化、恢复、持续高位和混合轨迹。病例的选择旨在呈现纵向轨迹模式的多样性，而非偏向某种特定的系统结局。参与者仅查看截至相应病例周可获得的信息；评估期间不会展示后续目标结局。

4.2.2 参与者。 我们招募了22名精神科临床医生，其中包括14名精神科住院医师和8名委员会认证精神科医生（女性7名；年龄 M=34.5, SD=6.4）。临床经验定义为培训和执业经验总年限，包括实习、住院医师培训及之后的临床实践。8名参与者有1–3年经验，6名有4–5年经验，3名有6–10年经验，5名有10年以上经验。完整的参与者人口统计资料见附录D。参与者获得约150美元的报酬。研究方案经机构审查委员会（IRB）批准，所有参与者均在参加研究前提供了知情同意。

4.2.3 实验流程。 在简短熟悉界面后，每位参与者审查了四份预先生成的患者病例报告，其中两份属于单智能体条件，另外两份属于 CHORAL 条件。为确保呈现一致性，报告采用统一格式，但未对其实质性推理和预后内容进行人工编辑。病例与条件分配以及呈现顺序在参与者之间进行了平衡，使每个病例在每种条件下均获得11次评估。每个病例结束后，参与者完成相应问卷；全部四个病例结束后，参与者参加半结构式访谈，比较两种条件。每次会话约持续40–50分钟，并在线进行。

4.2.4 测量指标。 问卷评估了对应 RQ2 和 RQ3 的两组结果。对于临床有用性、临床可操作性和工作流程适配度 （RQ2），问卷条目评估了预后判断的临床有用性、整合多个领域的有用性、对识别需要进一步确认的信息及确定随访问题优先级的支持，以及所呈现信息对临床工作流程的适用性。对于可争辩性与依赖支持 （RQ3），问卷条目评估了临床医生对以下方面支持程度的感知：将最终评估追溯至其支持性证据和领域特定的预后评估，识别不同领域之间的分歧或替代性解释，确定推理中需要进一步审查的部分，以及区分系统输出中可以依赖的部分和需要进一步核实的部分。完整问卷见附录E。

半结构化访谈探讨了这两个条件如何支持临床医生理解和审查预后判断。我们考察了信息和推理轨迹如何支持临床医生质疑

表2。预 测性能：CHORAL与参考 nce 基线的比较。

条件	平衡准确率	精确率	召回率	F1	恶化检测
多数基线	0.500	0.000	0.000	0.000	0/295 (0%)
持久性基线	0.672	**0.486**	0.478	0.482	0/295（0%）
单智能体基线	0.639	0.351	0.558	0.430	104/295（35%）
CHORAL	**0.673**	0.403	**0.612**	**0.484**	**118/295（40%）**

表3。消融 关于 CHORAL 架构组件的结果。恶化表示对低到高转变的检测。

条件	平衡准确率	精确率	召回率	F1	恶化检测
无编排器	0.657	0.353	**0.638**	0.454	**137/295（46%）**
无记忆	0.641	0.393	0.571	0.461	103/295（35%）
无按需检索	0.640	0.354	0.560	0.433	129/295（44%）
CHORAL	**0.673**	**0.403**	0.612	**0.484**	118/295（40%）

评估、识别信息缺口和随访优先事项，并确定哪些输出有用、繁琐或值得信任，以及哪些需要进一步核实。

4.2.5 分析。 由于问卷回答采用7点李克特量表收集，我们使用Wilcoxon符号秩检验比较两个条件。对于每位参与者，我们分别对其在每个条件下被分配的两个案例的评分取平均值，从而针对每个条目得到每个条件下一个参与者层面的得分。我们报告匹配对秩二列相关（r_{rb}）作为效应量。对于多重比较，在研究问题2和研究问题3的测量指标系列内分别使用Benjamini– Hochberg程序，对条目层面的 p-值进行校正，FDR为 $q = .05$。

两位研究人员对访谈记录进行了定性编码，并通过讨论反复比较和完善编码及反复出现的主题。

5 结果

5.1 研究问题1。架构贡献

5.1.1 预后性能。 CHORAL 的平衡准确率为 0.673，与持久性基线（0.672）相当，高于单智能体基线（0.639）和多数类基线（0.500）（表2）。相较于单智能体基线，差异为 0.034（95% 置信区间 [0.008，0.058]）；相较于持久性基线，差异为 0.001（95% 置信区间 [-0.021，0.026]）。根据构造，持久性基线未检测到任何低到高转变，而 CHORAL 检测到 118 个（40.0%），单智能体基线检测到 104 个（35.3%）。相较于持久性基线，CHORAL 的召回率更高（0.612 对 0.478），但精确率更低（0.403 对 0.486）。

消融实验显示了互补的权衡（表3）。移除按需检索或记忆后，平衡准确率分别降至 0.640 和 0.641。移除编排器后，精确率从 0.403 降至 0.353，但召回率从 0.612 升至 0.638。尽管无编排器和无按需检索的恶化敏感度更高（分别为 46% 和 44%，而完整系统为 40%），CHORAL 的警报精确率最高（29%），误报率最低（14.4%；表4）。因此，CHORAL 的贡献并非一味提高敏感性，而是实现了不同的运行权衡：与某些消融版本相比，完整系统检测出的恶化病例较少，但误报更少，整体平衡准确率也更高。

表4。当前低状态案例的警报级性能（$n = 2,\ 342$）。警报表示预测的未来高状态；由于持久性基线按构造不会发出任何警报，因此予以省略。误报率根据低状态到低状态的案例（$n = 2,\ 047$）计算。

条件	恶化敏感性	警报精确率	误报 Rate
单智能体基线	35%	22%	17.7%
无编排器	46%	24%	20.7%
无记忆	35%	25%	15.4%
无按需检索	44%	25%	19.1%
CHORAL	40%	**29%**	**14.4%**

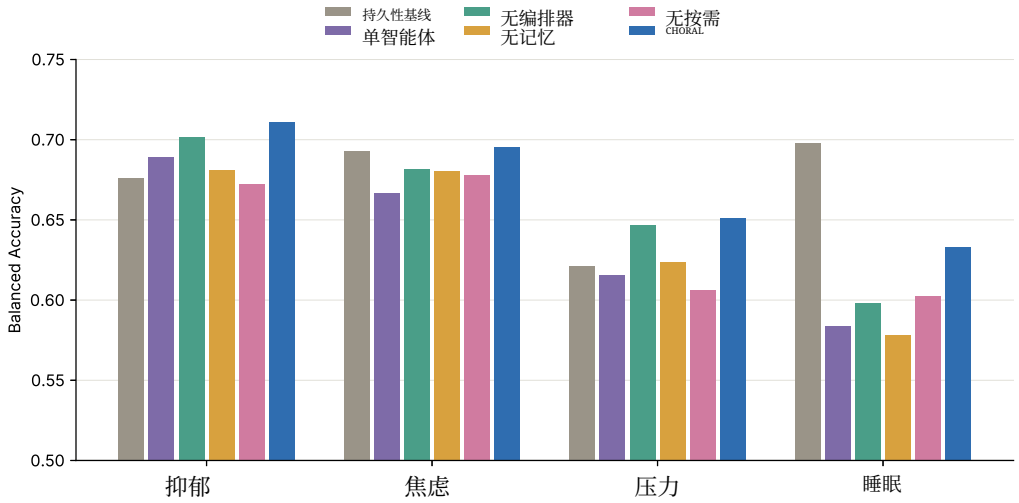


图 7. 持久性基线、单智能体基线、CHORAL 消融实验及完整 CHORAL 在心理健康领域的平衡准确率。

表 5. 单智能体基线、CHORAL 消融实验及完整 CHORAL 的证据依据。

条件	证据渠道覆盖率	无效证据引用率
单智能体基线	70.1%	0.0112%
无编排器	93.8%	0.0035%
无记忆	88.7%	0.0000%
无按需检索	87.4%	0.0000%
CHORAL	89.2%	0.0031%

在各领域中（图7），CHORAL 在抑郁（0.711）、焦虑（0.695）和压力（0.651）方面的平衡准确率最高。只有在睡眠领域，持久性基线的表现优于 CHORAL（0.698 对 0.633）；在该领域，高状态更常见（42%，而其他领域为 10–21%），且具有持久性：仅 34% 转为低状态，而其他领域为 54–60%。因此，CHORAL 较高的召回率和较低的特异度降低了睡眠领域的平衡准确率。恶化检测也因领域而异（21–58%），这表明汇总结果 40% 背后存在异质性。各领域的详细结果见附录 G.1。

表6. 临床医生对单智能体和 CHORAL 的临床有用性、可操作性及工作流程适配度评分。

问卷项目	单智能体 M（SD）	CHORAL M（SD）	Δ	W	p	q	r_{rb}
临床有用性							
1. 鉴于近期变化，具有临床实用性	4.75 (0.84)	5.07 (0.94)	+0.32	49.0	.103	.190	+0.43
2. 循证临床评价	4.75 (0.86)	**5.20 (0.84)**	**+0.45**	**11.0**	**.004**	**.028**	**+0.82**
3. 跨领域辅助解读	4.86 (1.00)	5.30 (1.09)	+0.43	40.5	.024	.083	+0.57
临床可操作性							
4. 识别出需要进一步核实的内容	4.57 (0.98)	4.84（1.17）	+0.27	27.5	.108	.190	+0.48
5. 预后问题的辅助优先级排序	4.95（0.95）	5.02（0.99）	+0.07	57.0	.561	.561	+0.16
工作流程适配度							
6. 信息量与组织	4.70 (1.10)	4.50 (1.18)	-0.20	52.5	.248	.347	-0.31
7. 可整合至现有工作流	4.57 (1.13)	4.70 (1.20)	+0.14	29.5	.449	.524	+0.24

5.1.2 证据使用与引文有效性。与单智能体基线相比，多智能体条件引用了范围更广的患者证据（表 5）。单智能体基线的证据渠道覆盖率为 70.1%，多智能体条件下为 87.4–93.8%，完整 CHORAL 达到 89.2%。不含编排器的条件下覆盖率最高（93.8%），这表明领域级评估引用的证据范围最广，而编排则产生了更具选择性的综合结果。

跨条件的无效证据引用率最高为 0.0112%。人工检查发现了两类不同的错误。在 CHORAL 和无 Orchestrator 条件下，智能体将 SOCIAL渠道缩写为 SOC；然而，被引用的社会暴露证据本身存在于患者记录中，且报告无误。在单智能体基线中，被引用的渠道和周数有效，但该记录中不存在具体的条目编号（例如，[UID179-SLEEP-23#10]，而 SLEEP渠道仅包含 #01–#04 条目）。因此，这些引用指向了不存在的证据条目，而 CHORAL 中的错误仍然指向患者记录中可识别的证据。

5.2 RQ2。临床有用性、临床可操作性和工作流程适配度

5.2.1 临床有用性。 在临床有用性方面，CHORAL 的总体评级高于单智能体。基于证据的临床评估显示出最大且在 Benjamini– Hochberg 校正后唯一具有统计显著性的差异（$\Delta = +0.45$、 $p = .004$、 $q = .028$、 $r_{rb} = .82$）；跨领域整合的评级也更高，但在校正后不再具有显著性（$\Delta = +0.43$、 $p = .024$、 $q = .083$）。

参与者表示，CHORAL 使最终预后判断依据的审查更加容易。尽管单智能体会呈现领域级发现和最终评估，但临床医生有时仍需推断这些发现如何促成最终判断。在 CHORAL 中，综合结果与审查过程使这种关联更加明确。P6 将单智能体条件下的这种情况描述为“中间存在的空隙”：“在单智能体版本中，我可以看到各项领域评估和最终结论，但两者之间仍然存在空隙。使用 CHORAL 时，我可以了解哪些问题被延续下来，哪些问题被弱化，以及解读在哪些地方发生了变化。”

参与者还重视能够查看最终摘要中未被强调的支持性发现。这些细节为他们审查最终评估提供了额外的情境信息，而不是仅仅依赖系统的主要结论。P15指出，“最终摘要聚焦于主要问题，但 CHORAL 仍然展示了那些未被强调的发现”，这帮助他们理解“哪些内容曾被考虑过，但被放在了次要位置”。P9同样强调，他们“无论如何都不会全盘接受人工智能的最终判断”，并且

“较小的线索和情境细节有时比结论本身更有用”，这有助于他们形成对患者的自身判断。

5.2.2 临床可操作性。在临床可操作性方面，CHORAL 的评分略高于单智能体，但两者的差异均无统计学意义。

参与者表示，当可用数据不足以让人轻易判断病例时，CHORAL 很有帮助，因为它能提供类似第二意见的另一种视角。P17 形容使用它“几乎就像征求第二意见”，而 P12 则指出，它促使自己重新考虑是否“忽略了对病例的另一种解读方式”。

与此同时，参与者指出，识别出需要进一步关注的问题，并不一定就能确定随访优先事项。优先级排序还取决于其他临床背景信息，包括患者的主诉、治疗和用药史，以及近期的生活事件。P19 指出：“系统可以告诉我哪些方面可能需要更多关注，但它确实无法决定什么应该优先处理。”P14 同样强调，即使两位患者的数据呈现出相似模式，医生首先关注的内容也可能不同，这取决于患者就诊的原因以及治疗过程中发生了什么。

参与者还发现，未解决的分歧有助于表明随访期间仍需核查哪些内容。P11 解释道：“分歧让我注意到预后中仍不确定的部分。它并不是让我直接接受最终评估，而是促使我去寻找更多情境信息，或向患者提出更具体的问题。”

然而，将这些已识别的关切转化为具体的随访问题仍然具有挑战性。尽管建议的问题有助于指出可能需要进一步探究的内容，但它们并不总是经过充分定制，无法直接用于与患者沟通。参与者指出，问题的措辞和表达方式需要根据具体患者和临床背景进行调整，例如患者的理解程度、与年龄相关的沟通偏好以及主要关切。P3 认为，这些问题有助于提醒自己需要涵盖哪些内容，但“相当笼统”，并指出会根据患者的不同而采用不同的措辞。P20 同样补充道，有些问题需要根据患者的情况进行简化或表达得更加直接。

5.2.3 工作流程适配度。 就工作流程适配度而言，两种条件之间的差异很小。CHORAL 在信息量与组织方面的评分略低（$\Delta = -0.20, q = .347$），而两种条件在融入现有工作流程方面的评分相近（$\Delta = +0.14, q = .524$）。

人们认为，CHORAL 在就诊前比就诊过程中更有用。P8 指出，它可以帮助自己“在进入诊室前快速了解发生了哪些变化，以及我应该关注什么”。

然而，参与者指出，就诊期间，信息量可能会变得令人不堪重负，因为临床医生本就需要关注电子病历、药物、既往记录以及与患者正在进行的对话。P14 表示，他们“可能不会从头到尾读完全部审议内容”；P2 也指出，额外增加几段推理内容，可能会让人难以找出重点。

因此，参与者更希望能够选择性地查看详细推理，而不是每次都审查完整的协调过程。他们尤其关注智能体意见不一致之处、评估经审查后发生变化之处，以及仍未解决的不确定性。P15 将这种偏好概括为希望了解“他们在哪些地方意见不一、发生了哪些变化，以及哪些问题仍未解决”。

5.3 研究问题3：可质疑性与依赖支持

研究问题3考察了临床医生对可质疑性与依赖支持的看法（表7）。

表7。临床医生对单智能体和 CHORAL 的可质疑性支持与依赖支持评分。

问卷项目	单智能体 M（SD）	CHORAL M（SD）	Δ	W	p	q	r_{rb}
可质疑性支持							
1. 支持证据的可追溯性	5.20 (0.57)	5.55 (0.79)	+0.34	45.0	.071	.086	+0.47
2. 识别 替代方案 解释 跨心理健康领域	4.25 (1.15)	**5.41 (0.91)**	**+1.16**	**7.5**	**.001**	**.006**	**+0.90**
3. 找出可据以提出质疑的依据	4.45 (1.18)	**5.07 (1.13)**	**+0.61**	**21.5**	**.003**	**.008**	**+0.77**
依赖与核验支持							
4. 作为临床参考的依赖	4.34（0.97）	4.68（1.01）	+0.34	33.0	.064	.086	+0.51
5. 区分信任与核实	4.61 (0.94)	**5.11 (0.79)**	**+0.50**	**24.0**	**.021**	**.042**	**+0.65**
6. 不当判断风险低	5.32 (0.80)	5.43 (0.99)	+0.11	17.0	.265	.265	+0.38

5.3.1 可质疑性支持。 与单智能体相比，CHORAL在帮助临床医生识别心理健康领域中的替代性解释（$\Delta = +1.16$, $p = .001$, $q = .006$,$r_{rb} = .90$），以及帮助他们找到质疑系统判断的依据方面（$\Delta = +0.61$, $p = .003$, $q = .008$, $r_{rb} = .77$），获得了显著更高的评分。

参与者指出，尽管单智能体呈现了领域级发现和最终判断，但他们仍然必须自行弄清评估中的哪些部分存在分歧，或希望进一步仔细检查哪些部分。P14 解释道：“在单智能体版本中，我可以看到最终判断，但如果我不同意，我仍然必须自行弄清楚，我究竟不同意推理中的哪一部分。”在 CHORAL 中，领域智能体提出的分歧和替代性解释仍然清晰可见，使临床医生更容易看出最终评估可能存在不同的解释。P4 指出：“当领域代理与编排器意见不一致时，我会开始质疑最终评估是否涵盖了全貌。在对预后感到放心之前，我希望知道是什么导致了这种差异。”

参与者还强调，分歧的价值来自其背后的推理和证据，而不仅仅是分歧本身。P1 解释道：“一致意见本身对我来说并没有那么重要。重要的是其背后的推理。如果某个智能体说‘我因为这种模式而同意’，或者说‘我不同意，因为这一发现指向了另一种方向’，这会为我的临床工作提供更多依据。”

随后，临床医生利用这些替代性解释来检查底层患者数据，并评估哪种解释得到了更充分的支持。P16 表示：“当我看到领域代理给出了不同的解释时，我回头查看底层患者数据，了解每个判断所依据的内容。这帮助我决定哪种解释得到了更充分的支持。”这也为临床医生质疑最终判断提供了更具体的依据。P18 将其描述为让他们有了“可以具体质疑的内容”，并解释道：“我不再只是想‘我不确定是否同意这个判断’，而是可以说，‘我认为这一解释对于我们现有的证据来说过于绝对。’”

5.3.2 依赖与核验支持。 在支持临床医生区分系统输出中哪些方面可以依赖、哪些方面需要进一步核实时，CHORAL 获得了显著更高的评分（$\Delta = +0.50$、 $p = .021$、 $q = .042$、 $r_{rb} = .65$）。在将系统作为临床参考方面，CHORAL 的评分也更高，但该差异未达到统计显著性（$\Delta = +0.34$、 $p = .064$、 $q = .086$）。

参与者表示，CHORAL 帮助他们决定系统推理的哪些部分可以依赖，哪些部分需要进一步核实。他们并非只是简单接受最终结论，而是将中间推理与自己的临床判断进行比较。P7 解释道：“看到中间推理并不会让我信任

它所说的一切。我可以将其与自己的想法进行比较，并决定哪些部分我可以放心依赖，哪些部分我仍然希望核实。”P17 同样指出：“我不会说自己对整个评估的信任程度都一样。”

参与者还指出，看到多种解释以及随后的协调过程，会影响他们对最终评估可信度的感受。P5 表示：“我更信任它了，因为系统看起来并不是直接跳到一个答案。我能看到，在得出最终评估之前，系统确实考虑过不同的解释。”P12 也强调，系统“没有掩盖不确定性”，并表明“在确定最终解释之前，同一种模式有几种可能的解读方式”。

即使没有查看每一步推理，一些参与者仍觉得，将 CHORAL 用作临床参考更令人安心，因为他们能看到最终评估经过了多角度审查。P10 总结道：“我可能不会读智能体之间的每一段交流，但只要知道最终答案经过了多角度核查，我就会更放心地使用它。”

6 讨论

6.1 纵向患者生成健康数据的多智能体推理

CHORAL 的平衡准确率与持久性基线相当，同时能够检测到 40% 已观察到的低状态到高状态的转变。由于持久性基线按其构造会延续当前状态，因此它无法预判这类转变，属于结构性局限，而非其整体表现均较差的证据。与持久性基线相比，CHORAL 的召回率更高，但精确率较低，这凸显了识别未来高状态与避免假阳性之间的权衡。与单智能体基线相比，CHORAL 的平衡准确率更高，证据使用也更广泛。在心理健康感知中，这种更广泛的证据使用尤为重要，因为行为信号往往需要结合情境进行解释，而不能直接映射到症状标签 [17]。相关研究同样指出，评估临床感知系统时，不应只看预测性能，还应考量它们为护理提供的额外信息和可操作洞见 [1]。从这一角度看，CHORAL 的结果表明，纳入更广泛的患者证据并在必要时寻求更多情境信息具有价值，而不应仅依赖预测性能。

消融结果进一步表明，CHORAL 增加的推理阶段并非简单冗余。移除按需检索后，平衡准确率、召回率和 F1 值降幅最大，这与在初始可用信息不足时补充证据有所帮助的情况一致。移除记忆也降低了平衡准确率和召回率，表明患者特异性历史语境发挥了一定作用。相比之下，移除编排器提高了召回率，却降低了精确率，同时也提高了证据渠道覆盖率。这一模式与以下观点一致：编排的作用较少在于扩大所考虑证据的数量，更多在于比较领域级判断，并抑制过强的高状态预测。综合来看，检索、记忆和编排发挥了互补作用：补充证据、提供历史背景以及调整判断。因此，CHORAL 的价值与其说在于使用多个智能体本身，不如说在于围绕互补的信息功能来构建纵向患者生成健康数据推理。这与先前有关专科角色和协调的医学多智能体研究 [11, 35, 71] 一致，也与健康数据意义建构研究相符；后者表明，纵向解读往往需要整合多个数据源，将当前观察结果与既往情境联系起来，并在观察到的模式无法得到充分解释时寻求更多信息 [53]。

然而，更丰富的推理并未在每个心理健康领域都带来最高性能。对于抑郁、焦虑和压力，CHORAL 的表现优于持久性基线，而在睡眠领域，持久性取得了更高的平衡

准确率。在我们的数据中，睡眠高状态相比其他心理健康领域的高状态，也更有可能在后续评估中保持为高状态，这表明近期的睡眠状态本身已经为未来状态提供了相对较强的信号。因此，额外推理的相对收益可能部分取决于目标状态随时间保持不变的程度，而所有领域未必都需要相同水平的多智能体推理。这一观察结果与根据任务复杂度调整协作结构的自适应多智能体方法一致 [35]。因此，未来的系统可以探索有选择地分配更深入的推理，以应对可能需要额外证据、历史背景或跨领域协调的情况。

6.2 面向临床医生审查的多智能体审议

在临床医生评估中，与单智能体条件相比，CHORAL在帮助临床医生识别替代性解释以及找到质疑系统判断的依据方面获得了显著更高的评分。两种条件之间最大的差异体现在临床医生对支持识别替代性解释程度的评分上。定性结果进一步表明，临床医生重视领域智能体评估与编排器综合判断之间的分歧，尤其是在这些分歧背后的推理和证据清晰可见时。这些分歧促使临床医生重新审视相应的患者证据、比较相互竞争的解释，并确定最终判断中需要进一步审查的具体部分。因此，CHORAL的审议提供了具体的审查切入点，使临床医生能够检查并质疑系统的预后判断。然而，由于临床医生研究比较的是完整的系统配置，这些差异不能单独归因于智能体组织方式，也可能反映了CHORAL的推理、检索、审查和呈现机制的综合影响。

这一发现拓展了既有的临床可解释人工智能研究，该研究表明，临床医生会使用解释，将人工智能输出与自己的判断进行比较，并验证其依据 [58, 63]。临床医生可能会有选择地接受、修改或暂缓采纳人工智能建议的某些方面，而不是将其作为一个整体来对待 [58]。关于可质疑人工智能的研究同样强调，应提供能够让用户审问和质疑人工智能决策的信息 [52]。CHORAL的替代性解释和智能体间分歧通过揭示判断可能需要受到质疑的具体位置及原因，支持了这一过程，而不是仅仅呈现一个综合结论。

在支持临床医生区分哪些内容可以依赖、哪些内容需要进一步核实方面，CHORAL获得了显著更高的评价。尽管在CHORAL条件下，将系统作为临床参考的依赖程度也获得了更高评价，但这一差异未达到统计显著性；对于不当判断风险的评价在不同条件之间也没有差异。这些发现表明，受试者认为系统在选择性依赖和核实方面提供了更大支持，但并不能证明临床医生对系统输出的实际依赖程度已根据其正确性进行了校准。这一解释与人机决策领域的相关研究一致：其目标并非最大化信任，而是支持适当依赖，即用户在适当情况下依赖人工智能，并在必要时对其进行核实或否决[10, 42]。健康人工智能领域的研究同样表明，信任可以根据具体建议和决策情境选择性地形成 [67]。与此同时，P10的陈述表明，即使没有进行详细检查，可见的审议过程也能带来 reassurance。由于CHORAL的智能体共享同一个骨干模型，因此不应将智能体之间的一致意见解读为独立的临床佐证。

这些发现表明，面向临床医生的多智能体系统应保留能够切实支持审查的审议内容，尤其是相互竞争的解释、具有重大影响的分歧以及未解决的不确定性。与其呈现每一次智能体交互，系统不如优先呈现有助于临床医生

判断哪些地方需要进一步审查或核实的推理事件。下一节将讨论如何呈现此类审议内容，同时避免造成过重的信息负担。

6.3 将多智能体推理融入临床护理

在临床医生评估中，对于患者证据是否支持临床评估这一项，CHORAL 获得了显著更高的评分；跨领域解读的评分也更高，但经校正后不再具有显著性。相比之下，在信息量与组织，或融入现有临床工作流程方面，未发现明显差异。定性研究结果同样表明，临床医生认为 CHORAL 详细的领域推理以及审查—修订过程有助于理解病例，同时也指出，在会诊期间审查完整的审议过程可能会带来沉重的信息负担。综合来看，这些发现表明，更丰富的推理与在临床工作流程中有效呈现推理，是两个不同的问题。这与以往临床可解释人工智能研究的发现一致：解释是否有用，不仅取决于解释是否可用，也取决于其是否适配临床医生的任务与工作流程 [58, 63]。近期研究也指出，更丰富的解释可能有助于提升理解，但同时会延长决策时间，或增加时间紧迫的工作流程中的认知负担 [27]。这些发现凸显了在面向临床医生的多智能体系统中区分推理深度 与呈现深度 的必要性。CHORAL 的多智能体流程可能会生成详细的中间推理，但并非所有推理都需要以同等显著的方式呈现。我们的定性研究结果表明，临床医生尤其关注那些会影响最终判断解读方式的推理事件，例如具有重大影响的分歧、修订后的解读以及未解决的不确定性。常规的一致意见或重复证据则可以压缩呈现，而在需要时仍可提供更详细的推理。这种分层呈现方式与以往关于临床人工智能选择性透明的研究一致；相关研究提出，可通过渐进式披露按需提供更多解释，同时管理信息负荷 [46]。因此，一个重要的设计问题不仅是系统进行多少推理，还包括应向临床医生展示其中哪些部分。

呈现推理的时机可能同样重要。参与者普遍认为，在就诊前审查 CHORAL 比在就诊过程中查看完整的推理过程更有价值。就诊前审查可以提供时间，以更深入地考察纵向轨迹、领域层面的解释、分歧和修订；而在就诊期间使用则可能需要快速获取关键问题、尚未解决的事项，以及仍需澄清的信息。因此，未来的系统可能需要根据工作流程阶段和临床医生当下的信息需求，调整推理的详细程度，而不是在所有临床环节采用固定的呈现方式。这与此前关于临床工作流程整合的研究一致；该研究强调，成功的整合不仅取决于将人工智能工具置于工作流程中，还取决于在适当的时间以适当的形式提供正确的信息 [55]。

最后，更好的解读并不一定会转化为更强的临床可操作性。尽管 CHORAL 提高了对系统输出中哪些部分需要核实的辨别支持，但并未明确提升临床医生识别患者病例中哪些方面仍无法通过可用数据得到解决，或确定随访中预后问题优先级的能力。定性研究结果进一步表明，确定随访优先事项通常需要纵向患者生成健康数据之外的临床背景信息，包括患者的主诉、用药和治疗史、近期生活事件以及持续存在的临床问题。参与者还指出，建议的随访问题在直接用于会诊之前，通常需要根据个体患者情况进一步调整。这些发现表明，解读纵向患者生成健康数据并将这种解读转化为临床行动，是相互关联但彼此 distinct 的步骤。心理健康感知领域的既往研究同样认为，感知所得的信息必须

与护理情境相联系，例如患者当前的问题、治疗目标和治疗反应，才能具备临床可操作性 [1]。因此，未来的系统可能需要将基于患者生成健康数据的多智能体推理与更广泛的临床病史及临床医生提供的背景信息相整合，而不是仅依赖患者生成健康数据来生成完整的随访计划。

6.4 局限性与未来工作

本研究存在若干局限性。首先，CHORAL 的定量评估采用了纵向大学生队列，而非在精神科门诊护理中收集的数据，这可能限制其对临床人群和门诊感知场景的普适性。其次，评估是在特定的大语言模型配置和系统实现下开展的，我们并未考察在其他基础模型或替代性多智能体架构中，是否也能取得类似的表现并产生类似的临床医生评估效果。第三，临床医生评估比较的是完整的系统配置，而非分别操控智能体组织方式、推理机制和呈现方式。因此，观察到的临床医生评估差异不应被解读为单由多智能体组织方式导致的因果效应。第四，临床医生评估并非在真实临床场景中开展，也没有让临床医生审查自己的患者；相反，参与者评估的是研究人员准备的病例。因此，未来研究应在真实临床场景中考察 CHORAL，使临床医生能够在持续开展患者护理的同时使用该系统。最后，CHORAL 主要依赖纵向患者生成健康数据，未纳入临床医生认为对随访决策很重要的更广泛临床背景信息，例如主诉、治疗史、用药信息和近期生活事件。未来研究应整合这些额外的临床背景信息，并考察它们能否改善预后推理，以及对核实和随访优先级排序的支持。

7 结论

本研究探讨了多智能体推理如何支持在心理健康随访中基于患者生成的健康数据开展纵向预后推理。我们开发了 CHORAL，这是一种多智能体脚手架，用于组织纵向轨迹解读、领域特定评估、信息缺口处理、跨领域综合和审查，同时确保底层证据和修订过程可供检查。在定量系统评估和临床医生研究中，与单智能体基线相比，CHORAL 展现出更高的预后性能和更广泛的证据使用；临床医生也认可它在提出替代性解释、质疑和选择性核实方面提供的支持。这些发现凸显了呈现具有实质影响的审议的价值，同时也表明，在时间受限的临床工作流程中，应将推理深度与呈现深度区分开来。

参照

[1] Daniel A Adler、Yuewen Yang、Thalia Viranda、Xuhai Xu、David C Mohr、Anna R Van Meter、Julia C Tartaglia、Nicholas C Jacobson、Fei Wang、Deborah Estrin 等。2024年。超越检测：迈向临床心理健康护理中可操作的传感研究。*ACM 交互、移动、可穿戴与泛在技术学报*8，4（2024年），1–33。[2] Kars Alfrink、Ianus Keller、Gerd Kortuem 和 Neelke Doorn。2023年。设计即具可质疑人工智能：迈向一种框架。《心智与机器》33，4（2023年），613–639。[3] Mohammad Alkhalaf、Ping Yu、Mengyang Yin 和 Chao Deng。2024年。应用结合检索增强生成的生成式人工智能，对电子健康记录中的关键信息进行总结与提取。《生物医学信息学杂志》156（2024年），104662。[4] 美国精神医学学会。2022年。《精神障碍诊断与统计手册》。美国精神医学学会出版社。[5] Brunilda Balliu、Chris Douglas、Darsol Seok、Liat Shenhav、Yue Wu、Doxa Chatzopoulou、William Kaiser、Victor Chen、Jennifer Kim、Sandeep Deverasetty 等。2024年。基于智能手机收集的行为模式进行个性化情绪预测。*NPJ Digital Medicine* 7，1（2024年），49。[6] Bei Bei、Joshua F Wiley、John Trinder 和 Rachel Manber。2016年。超越均值：关于睡眠／觉醒模式每日个体内变异性相关因素的系统性文献综述。*Sleep Medicine Reviews*28（2016年），108–124。

[7] Guanqun Bi、Zhuang Chen、Zhoufu Liu、Hongkai Wang、Xiyao Xiao、Yuqiang Xie、Wen Zhang、Yongkang Huang、Yuxuan Chen、Libiao Peng 等。2025年。MAGI：用于精神科评估的多智能体引导式访谈。载于《计算语言学协会论文集：ACL 2025年研究成果》。24898–24921。[8] Denny Borsboom。2017年。精神障碍网络理论。《世界精神病学》16，1（2017年），5–13。[9] Luca Canzian 和 Mirco Musolesi。2015年。抑郁症轨迹：通过分析智能手机移动轨迹对抑郁状态进行无干扰监测。载于《2015年ACM普适与泛在计算国际联合会议论文集》。1293–1304。[10] Shiye Cao、Anqi Liu 和 Chien-Ming Huang。2024年。为适当依赖而设计：人工智能不确定性呈现、初始用户决策和用户人口统计特征在人工智能辅助决策中的作用。《ACM 人机交互论文集》8，CSCW1（2024年），1–32。[11] Kai Chen、Xinfeng Li、Tianpei Yang、Hewei Wang、Wei Dong 和 Yang Gao。2025年。Mdteamgpt：用于多学科团队医疗会诊的自我演进型基于大语言模型的多智能体框架。arXiv预印本 arXiv:2503.13856（2025年）。[12] Adrien Choi、Aysel Ooi 和 Danielle Lottridge。2024年。压力、焦虑和轻度抑郁的数字表型分析：系统性文献综述。JMIR mHealth and uHealth 12，1（2024年），e40689。[13] Akshat Choube、Ha Le、Jiachen Li、Kaixin Ji、Vedant Das Swain 和 Varun Mishra。2025年。GLOSS：用于对健康与福祉被动传感数据进行开放式意义建构的大语言模型群。ACM《交互、移动、可穿戴与泛在技术论文集》9，3（2025年），1–32。[14] Diana E Clarke 和 Emily A Kuhl。2014年。DSM-5跨领域症状测量：迈向未来精神科护理的一步？《世界精神病学》13，3（2014年），314。[15] William Dormechele、Isaac Yeboah Addo、Caleb Boadi、Emmanuel Osei Bonsu 和 Mercy Oseiwah Adams。2026年。心理健康领域用于复发检测的数字表型分析与数字监测技术：系统性综述。BMC Psychiatry（2026年）。[16] Christopher S Douglas、Eliza Congdon、Crane Huang、Darsol Seok、Zachary D Cohen、Samir Akre、Veronica Tozzo、Feiyang Huang、Danielle Ramo-Larios、Raphe A Bernier 等。2026年。数字心理健康研究评估了大规模数字感知用于抑郁和焦虑研究的可行性。npj Digital Public Health 1，1（2026年），27。[17] Zachary Englhardt、Chengqian Ma、Margaret E Morris、Chun-Cheng Chang、Xuhai" Orson" Xu、Lianhui Qin、Daniel McDuff、Xin Liu、Shwetak Patel 和 Vikram Iyer。2024年。从分类到临床洞见：迈向利用大语言模型分析和推理移动与行为健康数据。ACM《交互、移动、可穿戴与泛在技术论文集》8，2（2024年），1–25。[18] Aaron J Fisher。2015年。迈向心理评估的动态模型：对个性化护理的启示。《咨询与临床心理学杂志》83，4（2015年），825。[19] John C Fortney、Jürgen Unützer、Glenda Wrenn、Jeffrey M Pyne、G Richard Smith、Michael Schoenbaum 和 Henry T Harbin。2017年。基于测量的护理迎来转折点。《精神科服务》68，2（2017年），179–188。[20] Ethan Goh、Robert Gallo、Jason Hom、Eric Strong、Yingjie Weng、Hannah Kerman、Joséphine A Cool、Zahir Kanjee、Andrew S Parsons、Neera Ahuja 等。2024年。大语言模型对诊断推理的影响：一项随机临床试验。JAMA Network Open 7，10（2024年），e2440969。[21] Adam Greene、Neviah Blair、Samin Mahdipour Aghabagher、Simmi Kumari、Michael W Schlund、Alex Fedorov、Vince D Calhoun、Xinhui Li 和 Rogers F Silva。2025年。人工智能精神科医生助手：一种基于大语言模型的多智能体系统，用于根据临床访谈进行抑郁症评估。载于《健康机器学习》2025年。[22] C Heneghan、P Glasziou、M Thompson、P Rose、J Balla、D Lasserson、C Scott 和 R Perera。2009年。初级保健中采用的诊断策略。BMJ 338（2009年）。[23] Joy Higgs、Gail M Jensen、Stephen Loftus、Franziska V Trede 和 Sandra Grace。2024年。《卫生专业临床推理》电子书。爱思唯尔健康科学。[24] Max Hirshkowitz、Kaitlyn Whiton、Steven M Albert、Cathy Alessi、Oliviero Bruni、Lydia DonCarlos、Nancy Hazen、John Herman、Eliot S Katz、Leila Kheirandish-Gozal 等。2015年。美国国家睡眠基金会睡眠时长建议：方法与结果摘要。《睡眠健康》1，1（2015年），40–43。[25] Ran Ha Hong、Jill K Murphy、Erin E Michalak、Trisha Chakrabarty、Zuowei Wang、Sagar V Parikh、Larry Culpepper、Lakshmi N Yatham、Raymond W Lam 和 Jun Chen。2021年。实施抑郁症基于测量的护理：针对精神科医生和初级保健医生的实用解决方案。《神经精神疾病与治疗》（2021年），79–90。[26] Tianyi Huang、Sara Mariani 和 Susan Redline。2020年。睡眠不规律与心血管事件风险：多族裔动脉粥样硬化研究。《美国心脏病学会杂志》75，9（2020年），991–999。[27] Tim Hunsicker、André Schulz、Robert Andreas Leist、Stephan Kiefer、Karl T Boden、Kai Rothaus、Cornelius J König 和 Markus Langer。2026年。可解释人工智能在临床诊断与治疗人机协作工作流程中的效率陷阱。《人为因素》（2026年），00187208261443764。[28] Jonathan S Ilgen、Kevin W Eva 和 Glenn Regehr。2016年。标签中包含什么？诊断是临床推理的起点还是终点？《普通内科医学杂志》31，4（2016年），435–437。doi:10.1007/s11606-016-3592-7[29] Thomas Insel、Bruce Cuthbert、Marjorie Garvey、Robert Heinssen、Daniel S Pine、Kevin Quinn、Charles Sanislow 和 Philip Wang。2010年。研究领域标准（RDoC）：迈向一种用于精神障碍研究的新分类框架。《美国精神病学杂志》167，7（2010年），748–751。[30] Thomas R Insel。2017年。数字表型分析：行为科学的新技术。《JAMA》318，13（2017年），1215–1216。[31] Yugyeong Jung、Thu Hoang Anh Vo、Hyun Seung Moon、Jae Young Choi、Hyangkyeong Oh、Ujin Lee、Eunjoo Kim、Tak Yeon Lee 和 Uichin Lee。2026年。迈向灵活的精神科病史采集与可视化：利用大语言模型探索临床医生观点。载于《2026年CHI计算机系统中的人为因素会议论文集》。1–23。

[32] Mohammad Amin Kamaleddin、Mina Mirjalili、Reza Barzegar、Nghia Trung Le、Areez Visram、Fatemeh Gholami Nezhad、Alice Rueda、Bazen Gashaw Teferra、Candice Monson、Tristan Glatard 等。2026。用于结构化临床访谈与精神科筛查的多智能体大语言模型框架：概念验证。《转化精神病学》（2026）。doi:10.1038/s41398-026-04335-5[33] Jerome P Kassirer。1983。通过迭代假设检验教授临床医学：让我们宣讲我们实践的内容。《新英格兰医学杂志》309，15（1983），921–923。[34] Ah Young Kim、Seonmin Kim、Jisu Lee、Youngwoong Han、Heon-Jeong Lee 和 Chul-Hyun Cho。2026。基于智能手机的数字表型分析，用于在韩国社区环境中检测高风险抑郁症和焦虑。《互联网干预》44（2026），100934。[35] Yubin Kim、Chanwoo Park、Hyewon Jeong、Yik S Chan、Xuhai Xu、Daniel McDuff、Hyeonhoon Lee、Marzyeh Ghassemi、Cynthia Breazeal 和 Hae W Park。2024。Mdagents：用于医疗决策的大语言模型自适应协作。《神经信息处理系统进展》37（2024），79410–79452。[36] Yubin Kim、Xuhai Xu、Daniel McDuff、Cynthia Breazeal 和 Hae Won Park。2024。Health-LLM：利用可穿戴传感器数据进行健康预测的大语言模型。载于《2024年健康、推理与学习会议论文集》（CHIL），第248卷。PMLR，522–539。[37] Kurt Kroenke、Robert L Spitzer 和 Janet BW Williams。2003。患者健康问卷-2：两项抑郁症筛查工具的有效性。《医疗护理》41，11（2003），1284–1292。[38] Kurt Kroenke、Robert L Spitzer、Janet BW Williams、Patrick O Monahan 和 Bernd Löwe。2007。初级保健中的焦虑障碍：患病率、功能损害、共病与检测。《内科学年鉴》146，5（2007），317–325。[39] Elizabeth W Lampe、Amanda C Collins、Ahhyun Lee、Nicholas Enbar-Salo、Tess Z Griffin、Arvind Pillai、Michael V Heinz、Damien Lekkas、Matthew D Nemesure、Daniel M Mackin 等。2026。重性抑郁障碍数字表型的个体间差异：一项使用智能手机和可穿戴传感器数据的被动感知研究。《行为研究与治疗》（2026），104986。[40] Xiao Li、Zhuo Chen、Jun Xia、Hongxin Xiang、Chao Wang、Wenjie Du 和 Yang Wang。2026。Med-SRAF：通过语义路由与智能体融合实现医学推理的多智能体框架。载于《计算语言学协会：ACL 2026论文集：研究成果》。38009–38029。[41] Henrietta Lyons、Eduardo Velloso 和 Tim Miller。2021。将可质疑性概念化：对质疑算法决策的观点。《ACM人机交互汇刊》5，CSCW1（2021），1–25。[42] Shuai Ma、Qiaoyi Chen、Xinru Wang、Chengbo Zheng、Zhenhui Peng、Ming Yin 和 Xiaojuan Ma。2025。迈向人类与人工智能协商：大语言模型赋能的协商式人工智能的设计与评估，以支持人工智能辅助决策。载于《2025年CHI计算机系统中的人为因素会议论文集》。1–23。[43] Bruce S McEwen。2008。应激激素对健康与疾病的中枢效应：理解压力和应激介质的保护作用与损害作用。《欧洲药理学杂志》583，2-3（2008），174–185。[44] Tim Miller。2019。人工智能中的解释：来自社会科学的见解。《人工智能》267（2019），1–38。[45] Stéphane Mouchabac、Ismael Conejero、Camille Lakhlifi、Ilyass Msellek、Leo Malandain、Vladimir Adrien、Florian Ferreri、Bruno Millet、Olivier Bonnot、Alexis Bourla 等。2021。改善精神病学临床决策：实施数字表型分析或可减轻患者和临床医生个体认知偏差的影响。《临床神经科学对话》23，1（2021），52–61。[46] Deepa Muralidhar、Rafik Belloum 和 Ashwin Ashok。2025。在人工智能临床诊断系统中通过渐进式披露实现选择性透明。《国际人机交互研究杂志》（2025），103591。[47] Subigya Nepal、Wenjun Liu、Arvind Pillai、Weichen Wang、Vlado Vojdanovski、Jeremy F Huckins、Courtney Rogers、Meghan L Meyer 和 Andrew T Campbell。2024。捕捉大学生活体验：一项为期四年的移动感知研究，探究大流行病期间大学生的心理健康、复原力与行为。《ACM交互、移动、可穿戴与泛在技术汇刊》8，1（2024），1–37。[48] Hung Nguyen、Alireza Rahimi、Veronica Whitford、Hélène Fournier、Irina Kondratova、René Richard 和 Hung Cao。2026。Heart2mind：利用可穿戴心电图监测器的人本可质疑精神障碍预测系统。《ACM医疗保健计算汇刊》（2026）。[49] Barry Nurcombe 和 Rollin M Gallagher。1986。精神病学临床流程：诊断与管理规划。CUP Archive。[50] 共识会议专家组、Nathaniel F Watson、M Safwan Badr、Gregory Belenky、Donald L Bliwise、Orfeu M Buxton、Daniel Buysse、David F Dinges、James Gangwisch、Michael A Grandner 等。2015。健康成人的推荐睡眠时长：美国睡眠医学学会与睡眠研究学会的联合共识声明。《临床睡眠医学杂志》11，6（2015），591–592。[51] Joon Sung Park、Joseph O’Brien、Carrie Jun Cai、Meredith Ringel Morris、Percy Liang 和 Michael S Bernstein。2023。生成式智能体：人类行为的交互式拟像。载于第36届ACM用户界面软件与技术研讨会论文集。1–22。[52] Thomas Ploug 和 Søren Holm。2020。可争议人工智能诊断的四个维度——以患者为中心的可解释人工智能方法。《医学中的人工智能》107（2020），101901。[53] Shriti Raj、Joyce M Lee、Ashley Garrity 和 Mark W Newman。2019。置于情境中的临床数据：面向个人健康数据解读的意义建构工具。《ACM交互、移动、可穿戴与泛在技术汇刊》3，1（2019），1–20。[54] Sohrab Saeb、Mi Zhang、Christopher J Karr、Stephen M Schueller、Marya E Corden、Konrad P Kording、David C Mohr 等。2015。与日常生活行为中抑郁症状严重程度相关的手机传感器指标：一项探索性研究。《医学互联网研究杂志》17，7（2015），e4273。[55] Megan E Salwei、Pascale Carayon、Peter LT Hoonakker、Ann Schoofs Hundt、Douglas Wiegmann、Michael Pulia 和 Brian W Patterson。2021。急诊科中基于人为因素的临床决策支持系统的工作流程整合分析。《应用人体工程学》97（2021），103498。[56] Michael J Sateia。2014。国际睡眠障碍分类。《胸部》146，5（2014），1387–1394。

[57] Noah Shinn、Federico Cassano、Ashwin Gopinath、KarthikNarasimhan 和 Shunyu Yao。2023。Reflexion：通过言语强化学习训练语言智能体。《神经信息处理系统进展》36（2023），8634–8652。[58] Venkatesh Sivaraman、Leigh A Bukowski、Joel Levin、Jeremy M Kahn 和 Adam Perer。2023。忽略、信任还是协商：理解临床医生对医疗保健中基于人工智能的治疗建议的接受度。载于《2023年CHI计算机系统中的人为因素会议论文集》。1–18。[59] Venkatesh Sivaraman、Eric Paul Mason、Mengfan Ellen Li、Jessica Tong、Andrew Joseph King、Jeremy M Kahn 和 Adam Perer。2026。智能推理线索：复杂决策中人工智能信息作用的框架与案例研究。载于《2026年CHI计算机系统中的人为因素会议论文集》。1–21。[60] Caitlin A Stamatis、Jonah Meyerhoff、Yixuan Meng、Zhi Chong Chris Lin、Young Min Cho、Tony Liu、Chris J Karr、Tingting Liu、Brenda L Curtis、Lyle H Ungar 等。2024。被动感知的智能手机特征对抑郁和焦虑症状预测的时间效用差异：一项纵向队列研究。《npj心理健康研究》3，1（2024），1。[61] Jinwen Tang、Qiming Guo、Wenbo Sun 和 Yi Shang。2025。用于长上下文心理健康评估的分层多专家框架。载于《2025年IEEE人工智能会议（CAI）》。IEEE，435–440。[62] Xiangru Tang、Anni Zou、Zhuosheng Zhang、Ziming Li、Yilun Zhao、Xingyao Zhang、Arman Cohan 和 Mark Gerstein。2024。Medagents：作为零样本医学推理协作者的大语言模型。载于《计算语言学协会：ACL 2024论文集：研究成果》。599–621。[63] Sana Tonekaboni、Shalmali Joshi、Melissa D McCradden 和 Anna Goldenberg。2019。临床医生的诉求：为临床终端使用构建可解释机器学习的情境。载于《健康机器学习会议》。PMLR，359–380。[64] John Torous、Keris Jän Myrick、Natali Rauseo-Ricupero 和 Joseph Firth。2020。数字心理健康与COVID-19：利用当下的技术，加速未来医疗服务可及性和质量的提升。《JMIR心理健康》7，3（2020），e18848。[65] Madhukar H Trivedi、A John Rush、Stephen R Wisniewski、Andrew A Nierenberg、Diane Warden、Louise Ritz、Grayson Norquist、Robert H Howland、Barry Lebowitz、Patrick J McGrath 等。2006。利用STAR*D中的基于测量的护理评估西酞普兰治疗抑郁症的效果：对临床实践的启示。《美国精神病学杂志》163，1（2006），28–40。[66] T Bedirhan Üstün。2010。测量健康与残疾：世界卫生组织残疾评定量表（WHODAS 2.0）手册。世界卫生组织。[67] Sterre Van Arum、Hüseyin Uğur Genç、Dennis Reidsma 和 Armağan Karahanoğlu。2025。选择性信任：理解个人健康决策过程中的人机智能伙伴关系。载于《2025年CHI计算机系统中的人为因素会议论文集》。1–21。[68] Edward H Wagner、Brian T Austin、Connie Davis、Mike Hindmarsh、Judith Schaefer 和 Amy Bonomi。2001。改善慢性病护理：将证据转化为行动。《卫生事务》20，6（2001），64–78。[69] Haochun Wang、Sendong Zhao、Zewen Qiang、Nuwa Xi、Bing Qin 和 Ting Liu。2024。超越直接诊断：基于大语言模型的多专科智能体会诊与自动诊断。arXiv预印本 arXiv:2401.16107（2024）。[70] Ziyue Wang、Junde Wu、Linghan Cai、Chang Low、Xihong Yang、Qiaxuan Li 和 Yueming Jin。2026。Medagent-pro：通过推理智能体工作流实现循证多模态医学诊断。载于《国际学习表征会议》，第2026卷。114360–114389。[71] Zixiang Wang、Yinghao Zhu、Huiya Zhao、Xiaochen Zheng、Dehao Sui、Tianlong Wang、Wen Tang、Yasha Wang、Ewen Harrison、Chengwei Pan 等。2025。Colacare：通过大语言模型驱动的多智能体协作增强电子健康记录建模。载于《2025年ACM网络会议论文集》。2250–2261。[72] Lawrence L Weed 等。1968。指导与教学型医疗记录。《新英格兰医学杂志》278，11（1968），593–600。[73] Yuqi Wu、Guangya Wan、Jingjing Li、Shengming Zhao、Lingfeng Ma、Tianyi Ye、Mike Zhang、Ion Pop、Yanbo Zhang 和 Jie Chen。2026。WiseMind：一种知识引导的多智能体框架，用于准确且富有同理心的精神科诊断。《NPJ Digital Medicine》（2026）。[74] Zongqi Xia、Prerna Chikersal、Shruthi Venkatesh、Elizabeth Walker、Anind K Dey 和 Mayank Goel。2025。利用被动感知行为和生态瞬时评估对多发性硬化症严重程度进行纵向数字表型分析：真实世界评估。《医学互联网研究杂志》27（2025），e70871。[75] Wenxuan Xu、Arvind Pillai、Subigya Nepal、Amanda C Collins、Daniel M Mackin、Michael V Heinz、Tess Z Griffin、Nicholas C Jacobson 和 Andrew Campbell。2026。LENS：通过将多模态感知与语言模型对齐，为心理健康生成大语言模型赋能的叙事综合。载于《计算语言学协会第64届年会论文集（第1卷：长篇论文）》。17458–17481。[76] Yuting Zhan、Hongwei Liu 和 Yu Wang。2025。抑郁症的数字表型分析：一种用于识别治疗反应新型行为和生理标志物的多模态被动感知方法。《精神病学研究杂志》（2025）。[77] Panyu Zhang、Gyuwon Jung、Jumabek Alikhanov、Uzair Ahmed 和 Uichin Lee。2024。一种基于移动传感器数据、可复现的压力预测流程。《ACM交互、移动、可穿戴与泛在技术汇刊》8，3（2024），1–35。[78] Yuwei Zhang、Kumar Ayush、Siyuan Qiao、A Ali Heydari、Girish Narayanswamy、Maxwell A Xu、Ahmed A Metwally、Shawn Xu、Jake Garrison、Xuhai Xu 等。2025。Sensorlm：学习可穿戴传感器的语言。arXiv预印本 arXiv:2506.09108（2025）。[79] Zeyu Zhang、Quanyu Dai、Xiaohe Bo、Chen Ma、Rui Li、Xu Chen、Jieming Zhu、Zhenhua Dong 和 Ji-Rong Wen。2025。基于大语言模型的智能体记忆机制综述。《ACM信息系统汇刊》43，6（2025），1–47。[80] Zelun Tony Zhang、Sebastian S Feger、Lucas Dullenkopf、Rulu Liao、Lou Süsslin、Yuanting Liu 和 Andreas Butz。2024。超越建议：从对飞行员决策过程的事后支持到事前人工智能支持。《ACM人机交互汇刊》8，CSCW2（2024），1–32。[81] Wanjun Zhong、Lianghong Guo、Qiqi Gao、He Ye 和 Yanlin Wang。2024。Memorybank：通过长期记忆增强大语言模型。载于《AAAI人工智能会议论文集》，第38卷。19724–19731。

[82] Yinghao Zhu、Dehao Sui、Zixiang Wang、Xuning Hu、Lei Gu、Yifan Qi、Tianchen Wu、Ling Wang、Yuan Wei、Wen Tang等。2026年。通过交互式、可解释的人工智能副驾驶辅助临床决策：一项在肾脏病学和产科学领域与临床医生开展的真实世界用户研究。发表于*2026*年计算机系统中的人为因素会议论文集。1–26。
[83] Ruishi Zou、Shiyu Xu、MargaretE Morris、Jihan Ryu、Timothy D Becker、Nicholas Allen、Anne Marie Albano、Randy Auerbach、Dan Adler、Varun Mishra等。2026年。MIND：通过叙事式仪表板赋能心理健康临床医生，提供多模态数据洞察。发表于*2026*年计算机系统中的人为因素会议集。1–26。

附录

基于传感器的睡眠状态标签推导

由于CES不包含每周自我报告睡眠测量，我们根据每个7天窗口内的三个维度推导出二元睡眠状态：睡眠时长（平均<6 或 >10小时）、睡眠时间（均值起始时间早于21:00或晚于02:00）以及睡眠不规律（起始或结束时间标准差 >120 分钟）。这些研究特定的操作性界限参考了既往关于睡眠时长、睡眠-觉醒时间和个体内变异性的研究 [6, 24, 26, 50, 56] ，旨在将显著的睡眠模式偏离操作化，而非用于界定临床睡眠障碍阈值。在CES中，睡眠开始时间和结束时间以8分钟为单位编码，表示自20:00起经过的时间，因此跨午夜的睡眠时段仍保持连续尺度，不会发生时钟时间环绕。一周内三个维度中至少有两个被标记时，该周被标记为高，否则标记为低；这一聚合规则旨在捕捉多维度的睡眠模式偏离，而非临床睡眠障碍严重程度。

B CES特征到行为标记的映射

表 8 将 CES数据集中的被动感知特征映射到更高层级的行为标记及相应的、依据 WHODAS 确定的功能相关领域。为支持具有临床可解释性的推理，我们将低层级传感器信号归类为症状相关行为构念（例如，睡眠变化、注意力、疲劳）。这些构念进一步依据世界卫生组织残疾评定量表（WHODAS） [66], 与依据 WHODAS 确定的功能相关领域相对应，从而支持以症状和功能为中心、符合临床实践的解读。这些映射参考了数字表型分析和行为信号解读领域的既往研究；相关研究将被动感知特征与睡眠不规律、活动减少和社交退缩等临床相关构念联系起来 [17, 77]。此映射并不旨在暗示直接的因果关系，而是为与临床推理相关的行为模式提供可解释的代理指标。患者代理利用此映射，将其输出围绕症状-功能领域进行组织，而非围绕原始传感器数据流，从而使下游领域智能体能够基于结构化行为线索进行推理。

C 大语言模型实现与提示模板

CHORAL 采用 OpenAI API 实现为多智能体大语言模型系统。所有患者、领域和编排智能体均使用 GPT-5-mini 作为骨干模型。该模型无法配置温度参数，因此未对其进行明确指定。我们将每次智能体调用的 {max_completion_tokens} 设置为 16,000，且没有任何生成结果达到相应上限。然而，不同基于大语言模型的条件下的总计算量仍存在差异：从单智能体基线的每个病例均值 3.6 次调用和 77K 个总词元，到 CHORAL 的 16.1 次调用和 392K 个词元不等。

本附录总结了 CHORAL 和单智能体基线中使用的操作性提示模板。这些模板保留了智能体角色、任务指令、运行时输入、相关背景上下文、推理约束和结构化输出，同时为简洁起见省略了非操作性措辞和特定于实现的序列化内容。运行时变量用大括号表示（例如，{patient_evidence}），并在执行期间实例化。预后

表 8. CES 被动感知特征与症状或行为标记及 WHODAS 指导的功能相关领域的概念映射。该映射用于识别可能相关的行为线索，而非推断 WHODAS 分数或判定功能损害。

相关特征（来自 CES）	**症状／行为标记**	**WHODAS 指导的功能相关领域**
睡眠_时长，睡眠_结束时间，睡眠_开始时间	睡眠变化	生活活动；自我照护
loc_study_unlock_时长，loc_study_unlock_数量，loc_home_unlock_时长，loc_home_unlock_数量，解锁_num_小时_*，其他_游玩_num_小时_*，解锁_num_集_0，其他_游玩_num_集_0	注意力减退／难以集中注意力	认知；生活活动
loc_社交_持续时间，loc_其他_宿舍_持续时间，loc_休闲_持续时间，地点_食物_时长，其他_玩耍_时长_事件_0，其他_玩耍_次数_事件_0，步数_事件_0，地点_距离_事件_0，活动_骑自行车_事件_，活动_跑步_事件_0，活动_步行_事件_0，地点_锻炼_时长，地点_家中_时长，地点_学习_时长	兴趣／乐趣	与他人相处；社会参与 社会参与；自我照护
步数_周期_0，LOC_距离_周期_0，活动_骑行_自行车_周期_0，活动_跑步_周期_0，活动_步行_周期_0，步数_小时_*，LOC_距离_小时_*，LOC_锻炼时长，LOC_就诊_次数_周期_0	疲劳 / 精力不足	活动能力；认知
loc_social_dur，loc_other_dorm_dur，loc_就诊_num_ep_0	社交模式变化	与他人相处；社会参与 社会
解锁_时长_集_0，解锁_数量_集_0，其他_播放_次数_次_0，其他_播放_时长_次_0，解锁 _时长 _hr_*，解锁 _num_hr_*，其他_播放_时长_小时_*，loc_居家_解锁_时长	手机使用模式变化（尤其是夜间）	生活活动；认知

时间范围在我们的实验中被实例化为提前两个 EMA 步骤（约两周）。部分提示会根据条件作为推理工作流的一部分被调用。完整的已部署提示见补充材料。

C.1 患者代理：轨迹证据

在进行特定领域的预后评估之前，患者代理会根据患者生成健康数据构建一条共享的、领域无关的纵向轨迹。

患者代理：轨迹证据角色。 你是一个患者代理，负责根据由传感器数据和自我报告数据构建的表型快照，整理纵向行为模式。你的职责是为下游领域智能体构建共享的患者层面轨迹，而不是作出特定领域的预后判断。**任务**。 通过以下方式总结患者的纵向行为轨迹：1. 使用统计、结构和时间特征描述当前行为状态；2. 将当前状态分别与近期基线和稳定基线进行比较；3. 识别表明当前行为模式可能持续的持久性信号；4. 识别表明恶化、恢复或变化混合的变化信号；5. 比较可用观察窗口中的模式，以识别复发、持久性或偏离患者既往病程的情况；6. 在现有证据支持时，呈现与功能相关的行为背景。**输入**。 包含{observation_windows}期间行为特征的表型快照；{recent_baseline}；{stable_baseline}；来自表型数据库的纵向记录；以及{evidence_catalog}。**背景／情境**。 行为特征可能包括统计特征、随时间发生的结构变化，以及一天中不同时段的时间模式。{evidence_catalog}中的证据标识符可唯一标识患者层面证据，应在下游推理中予以保留。**规则**。 每一条患者特异性陈述都必须以相应的证据标识符为依据。当时间趋势与相对于基线的偏离不一致时，必须明确保留这一冲突。如果关于持久性或方向性变化的证据较弱或相互冲突，应如实说明，不要强行作出单一解释。不得诊断心理健康状况、作出特定领域的预后判断，或推断功能损害。

输出。 行为摘要；基线比较；持久性信号[]；变化信号[{direction}]；与功能相关的信号[]；历史背景[]；所引用的证据标识符[]。

C.2 领域代理：初步预后评估

分别对应抑郁、焦虑、压力和睡眠的四个领域智能体，独立评估同一条共享患者轨迹。

领域代理：初步预后评估

角色。 你是一名负责 {domain} 的临床专科医生。你的职责是从所分配领域的角度，评估患者在未来 {prediction_horizon} 内针对 {target_variable} 的可能症状变化过程。**任务**。 执行领域特定预后评估，具体包括：1. 识别与 {domain} 相关的纵向证据，包括趋势、复发、不稳定性、相对于基线的偏离，以及持久性或变化；2. 在患者轨迹支持的情况下，描述与功能相关的行为背景；3. 评估相关行为模式在预测时限内是否可能持续、恶化、恢复或呈现混合变化；4. 推断由此产生的症状病程风险，并从 {label_set} 中指定一个标签；5. 在相关时，利用检索到的患者特异性记忆，将当前轨迹与早前的评估及其后观察到的症状状态进行比较；6. 识别支持或削弱当前评估的证据；7. 识别尚未解决的信息需求以及相应的随访或管理考量。**输入**。患者代理轨迹；{retrieved_memory}，其中包含最近的 {k_memory} 条已确认领域记忆记录（若有）；以及与患者代理轨迹相关联的证据标识符。每条已确认记忆均包含早前的患者轨迹、先前的领域特定评估、随后观察到的症状状态，以及一段反思，说明早前评估及其支持性证据与后续病程相符或不同之处。每条检索到的记忆记录都可能提供可引用的既往案例标识符 [PC-. . .]。**背景 / 情境**。{reference_information}；用于界定目标类别的 {label_cutoffs}；包含基于 DSM 的标准或相关压力框架的 {domain_knowledge}；以及 {ema_variable_definitions}。**规则**。 这是一个预后任务，关注观测窗口之后的症状变化过程，而非诊断或单纯描述当前状态。优先考虑趋势、复发、不稳定性和相对于基线的变化等纵向信号，而非孤立观察。兼顾支持性证据和削弱性证据。从所分配领域的角度解读共享轨迹，而不要采用其他领域的解读。不得仅凭传感证据推断功能损害。引用患者特异性证据时使用其证据标识符；若检索到的记忆对评估有所贡献，则引用 [PC-. . .]。将评估视为暂定结论，并从 {label_set} 中选定一个标签。默认以当前观察到的状态作为结果；只有在引用至少一个升级或降级信号及其证据标识符时，才偏离该状态。**输出**。 预测标签；推理{证据轨迹[]， 与功能相关的背景、预后展望、症状风险推断}；支持性证据[]；削弱性证据[]；不确定性来源[]；信息需求[]；随访问题；干预提示；引用的证据标识符[]。

C.3 领域记忆

每轮推理后，领域代理会将其领域特定评估作为待确认的记忆条目存储。一旦相应的未来目标状态可用，系统便会将先前的评估与观察到的病程进行比较，并生成反思，说明先前的解读在哪些方面与后续结果一致或存在差异、哪些证据仍具有信息价值，以及哪些解读内容得到的支持较弱。由此产生的条目将成为已确认记忆，可在后续推理轮次中检索。在我们的实现中，每个领域代理会接收最近的三条已确认记忆。当先前的预测与观察到的状态不一致时，系统会撰写反思；对于正确的预测，则仅根据结果等级予以确认。此外，每个领域代理还会接收针对该患者自身已确认记忆的汇总统计，其中涵盖过去偏离当前状态的结果，以及先于已确认状态转变出现的信号。完整的记忆构建和反思提示见补充材料。

C.4 信息缺口处理

当领域代理识别出可能有助于澄清其初始评估的未解决信息时，CHORAL 会区分可从现有纵向记录中检索的信息与需要后续临床随访的信息。内部信息需求会触发针对性的患者智能体检索，随后进行一次领域智能体细化；而外部信息需求则保留为临床医生的随访问题。当初始评估偏离当前观察到的状态时，将调用此步骤。

C.4.1 领域代理：信息缺口查询。

领域代理：信息缺口查询

角色。 你是一名{domain}领域的临床专科医生，正在审查当前预后评估中尚未解决的信息需求。**任务**。 确定哪些额外信息最有助于明确当前评估，并将每项信息需求归类为可通过内部或外部途径解决。对于**内部缺口**，拟定有针对性的系统级查询，以便从现有纵向患者记录中获取答案。对于**外部信息缺口**，针对无法从现有患者生成健康数据或自我报告记录中获取的信息，拟定临床医生级随访问题。对于每项信息需求，说明其有助于明确的预期病程，以及它主要用于指导干预还是监测。**输入**。 领域代理的评估，包括{uncertainty_sources}和已识别的信息需求；患者代理报告；纵向患者记录中可用的历史背景；以及可从现有患者数据中查询的{allowed_signals}。**背景／情境**。 内部缺口可能涉及当前行为模式过去是否出现过、类似模式是否持续或已缓解，或该模式过去与症状状态如何共同出现。外部信息缺口涉及现有患者数据中未体现的信息，例如情境背景、主观体验，或患者对行为变化的解释。**规则**。 若指定了数量，则恰好生成{n_system}个系统随访问题和{n_clinician}个临床医生随访问题。系统随访问题必须请求可从现有纵向记录中检索的证据，而非要求对患者的完整病史进行宽泛的重新总结。临床医生随访问题必须涉及无法从现有记录中可靠推断的信息。在相关情况下，优先选择有助于明确复发、持久性、转变或恢复的查询。为每个项目标注其探查的预期病程（恶化、持久性基线或恢复），以及其指导的计划目的（干预或监测）。**输出**。 system_followups[{target_signal, anticipatory_focus, planning_type, purpose}]; clinician_followups[{question, anticipatory_focus, planning_type, purpose}]。

C.4.2 患者代理：定向证据检索。对于可在内部解决的信息需求，患者代理会从患者的纵向记录中有选择地检索更多证据，而不是重建完整的病程轨迹。与其他步骤不同，这种检索是通过对 PhenotypeDB 执行确定性程序实现的，而非调用大语言模型，因此返回的证据仅限于患者记录中已有的值。

患者代理：定向证据检索（程序）

输入。领域代理的系统级查询，包括从患者记录中可用信号里选定的 {target_signal} 及其指定用途。**流程**。对于每个系统级查询：1. 检索目标信号在当前窗口中的状态（窗口统计值、趋势，以及相对于近期基线和稳定基线的偏离）；2. 从纵向记录中检索此前最多五个观察窗口内的同一信号，并为每个窗口分配一个可引用的既往案例标识符 [PC-...]；3. 确定该信号何时偏离患者的基线，以及这种偏离已持续了多少个连续窗口（起始）；4. 识别该信号出现可比偏离的较早窗口，并报告该领域的自我报告症状状态在随后窗口中的发展情况（先例）；5. 总结与功能相关的行为信号相对于基线的变化（功能背景）。**规则**。仅使用截至当前推理轮次可获得的信息。不作领域特异性的预后判断。如果不存在所请求信号的当前证据或历史证据，则将查询返回为未解决。**输出**。target_signal；status（answered | unavailable）；current_signal；prior_report_context[{date, past_case_ref, signal_state, baseline_deviation}]; probes{onset, precedent, functional_impact}。

C.4.3 领域代理：评估修订。患者代理返回的证据会传回相应的领域代理，用于对其初始评估进行一次修订。

领域代理：评估修订

角色。你是一名 {domain} 临床专科医生，在获得额外的患者特异性证据后，完善你的初步预后评估。**任务**。使用检索到的证据重新考虑初始评估。确定额外证据是否支持保留或强化初始解读，是否表明应修订该解读，或是否使相关不确定性仍未解决。**输入**。初始的患者代理轨迹；领域代理的初始评估；系统级信息需求；患者代理检索到的证据；以及在可用时的 {retrieved_memory}。**规则**。将初始评估视为暂定评估。仅在检索到的证据与已确定的信息需求相关时使用该证据。同时考虑支持初始解读的证据和削弱初始解读的证据。如果额外证据不足，则保留未解决的不确定性。临床医生级后续问题在此阶段不视为患者证据。不得引入患者代理未提供的患者信息。保留初始评估中的相关引文，并使用其证据标识符引用新检索到的患者证据。**输出**。与领域代理初始评估相同的结构化输出，并更新 predicted_label、reasoning、supporting_evidence[], weakening_evidence[],uncertainty_sources[], information_needs[], 以及 cited_evidence_ids[]。

C.5 协调器：跨领域综合

编排 器将四项领域特定评估整合为一项临时综合评估，同时保留
其支持 证据以及跨领域的差异。

编排器：跨领域综合

角色。 你是编排器，负责协调各专科领域智能体的领域特定预后评估，在 {prediction_horizon} 内形成综合评估。**任务**。 通过以下方式构建临时跨领域综合结果：1. 从每个领域代理中提取主要预后问题和支持性证据；2. 比较各领域特定评估之间的一致意见和差异；3. 识别得到多个领域视角支持的行为模式或预后问题；4. 保留仍局限于单一领域或在不同领域之间存在差异的解释；5. 分析一个领域中的证据或问题如何为另一个领域中的解释提供情境或调节其解读；6. 在相关时使用检索到的编排器记忆，以考虑此前相似的跨领域模式如何在同一患者身上发展；7. 将各领域特定评估整合为临时病例级评估，同时保留其基础证据及来源领域；8. 根据综合结果生成领域专属主张，供相应的领域智能体后续审查。**输入**。 {n_domains} 个领域代理的评估，包括其预后标签、支持性和削弱性证据、未解决的信息需求以及证据标识符；以及 {retrieved_memory}，在可用时由最近的 {k_memory} 条已确认的编排器记忆条目组成。每条已确认的编排器记忆都包含此前的最终综合评估、随后观察到的症状状态，以及对先前跨领域解读与后续病程之间关系的反思。**规则**。 不要仅为保持一致而将相互冲突的领域解释压缩为一致意见。保留有意义的差异和未解决的不确定性。保留每个领域自身的解释及其证据；不得用一个领域的评估替代另一个领域的评估。对领域层面解释的任何修改都必须以患者特异性证据为依据，并可追溯至领域代理的原始评估。不得臆造新的患者证据。对于实质性的患者特异性主张，保留支持性证据标识符。编排器记录每个领域的标签、其佐证情况以及任何修订；只有在有引用的升级信号或降级信号支持时，才可更改领域标签，绝不能因为其他领域为高或低而更改。每个领域主张都应指出一种具体解释，以便相应的领域智能体随后进行审查，而不应只是简单重述其预后标签。**输出**。
provisional_predictions；label_review；domain_appraisals；integrated_case_formulation；agreements[]；differences[]；cross_domain_links[]；domain_claims；remaining_uncertainty[]；follow_up_needs[]；cited_evidence_ids[]。

C.6 领域代理：审查

每个领域代理都会审查编排器初步综合结果中与自身领域相关的部分，并指出哪些解释应予以保留、限定或修订。为便于面向临床医生呈现，我们将主张级审查结论归为两类：对于已确认，归为“同意”；对于部分成立或已排除，归为“不同意”。因此，“不同意”表示某项主张需要加以限定、修订或删除，并不一定意味着完全否定。

领域代理：审查

角色。 你是一名{domain}领域的临床专科医生，正从所分配领域的角度审查编排器的初步综合评估。**任务**。 评估{own_domain_claims}中的每项主张，并判断编排器是否准确呈现了你原本的领域特定评估及其支持性证据。请为每项主张指定以下判定之一：**已确认**：现有领域专属证据适当地支持该主张。**部分成立**：总体解释有证据支持，但其中某些部分表述过度、支持不足或需要限定。**已排除**：领域专属证据与该主张相矛盾，或该主张基于对该证据的重大误解。另请指出是否存在以下情况：1. 你原始评估中的重要支持性证据或削弱性证据被遗漏；2. 你的领域特定解释被过度简化，或与其他领域的解释合并，导致重要区别消失；3. 相关的不确定性或后续跟进需求被遗漏；或4. 你的领域证据与综合结果中的另一领域特定解释直接冲突。**输入**。 你原始的领域代理评估及支持性证据；编排器的初步综合评估；以及{own_domain_claims}，其中包含综合结果中涉及你所分配领域的主张。**规则**。 请从所分配领域的角度评估主张。不要编造患者证据。确认、限定或质疑某项解释时，请使用现有的证据标识符。若证据支持领域间存在差异，请保留这些领域特定差异，不要强行要求各智能体达成一致。**输出**。 primary_evaluation{claims_evaluated[{verdict, supporting_evidence, counter_evidence}]}; omission_or_misrepresentation[]; cross_domain_notes[{target_domain, conflict_note, severity}]。

C.7 编排器：修订

编排器使用领域智能体的审查反馈修订初步综合结果。综合—审查—修订过程最多可重复三轮。

编排器：修订**角色**。 你是编排器，负责在领域智能体审查后修订现有的综合评估。**任务**。 通过以下方式修订综合评估：1. 检查针对每项领域专属主张的审查反馈；2. 保留相应领域代理已确认的解释；3. 细化或限定收到部分成立判定的主张；4. 当收到已排除判定的主张与相应领域证据不一致时，删除或修订这些主张；5. 恢复被遗漏或错误表述的重要领域特异性发现、不确定性或信息需求；6. 纳入会实质性影响综合病例构建的跨领域冲突；以及7. 根据审查反馈和支持性证据，在必要时更新领域层面或综合结论。**输入**。 当前综合评估及其 {domain_claims}；领域代理的原始评估和支持性证据；领域智能体的主张层面评估；{omission_or_misrepresentation}反馈；以及{cross_domain_notes}。**背景／情境**。 修订后的综合结果可能会返回给领域智能体进行再次审查。综合—审查—修订过程最多重复{max_rounds}轮；在我们的实现中，{max_rounds}设为三轮。**规则**。 不得引入新的患者证据。修订必须以领域代理的原始评估、患者证据和领域代理的审查反馈为依据。除非额外的审查反馈提供了需要重新考虑的证据，否则应保留已确认的主张。当现有证据不支持调和时，应保留领域之间有意义的分歧。标签变更须遵循与综合阶段相同的证据条件。若不再需要进一步的实质性修订，过程可在{max_rounds}轮之前终止。**输出**。 更新后的 provisional_predictions；domain_appraisals；integrated_case_formulation；agreements[]；differences[]；cross_domain_links[]；domain_claims；remaining_uncertainty[]；follow_up_needs[]；以及cited_evidence_ids[]。

C.8 编排器记忆

审查—修订过程完成后，最终的综合评估会作为待处理的编排器记忆条目存储。相应的未来目标状态一旦可用，便会将先前的综合评估与观察到的跨领域进程进行比较，并生成反思。由此形成的已确认记忆会保留先前的综合解读、后续目标状态和结果反思，以供后续综合使用。

在我们的实现中，编排器会接收最近的三条已确认记忆。完整的记忆构建和反思提示见补充材料。

C.9 面向临床医生的最终报告

经过审查—修订过程后，编排器将完成的综合评估转化为结构化报告，并通过临床仪表板呈现。报告优先呈现患者近期的纵向轨迹和最终的领域级预后评估，随后列出相应的支持性证据、重要的跨领域一致之处或差异、剩余不确定性、随访问题以及干预或管理考量。报告阶段被要求不得引入新的患者证据，也不得自行更改已完成的综合评估。完整的报告提示见补充材料。

C.10 单智能体基线

单智能体基线保留了与 CHORAL 相同的患者代理轨迹、领域特定知识、目标定义、输出要求和可比的纵向记忆。它移除了 CHORAL 将任务拆分为独立领域代理的做法、按需进行的患者智能体检索和领域智能体细化过程，以及由协调器主导的综合、审查和修订。患者代理轨迹阶段予以保留，并使用与 CHORAL 相同的模板。随后，单个 GPT-5-mini 实例在一次推理过程中完成了四项领域特定评估和跨领域整合。

单智能体基线

角色。 你是一名单智能体临床推理智能体，负责预测患者在所有{n_domains}个领域中的症状变化过程，预测时限为{prediction_horizon}。你需要自行完成领域特定评估和跨领域整合，不得将任务分配给不同的专科代理。**任务**。

1. **逐领域评估**。 根据共享的患者轨迹，分别评估每个领域。对于每个领域，识别纵向支持性证据和削弱性证据、与功能相关的背景、预测时限内预期的症状变化过程、相关的患者个体病史、未解决的信息需求，以及随访或干预考量。确保每个领域具有相当的推理深度。

2. **跨领域综合**。 通过审视各领域之间共享的证据、一致意见、差异及相关关系，整合领域特定评估。当现有证据不足以支持统一解释时，应保留有意义的差异。
3. **最终确定**。 为每个领域生成一个预后标签，并给出病例层面的综合评估。**输入**。 {patient_evidence}，由 CHORAL 使用的同一患者代理轨迹构成，包括基线比较、纵向轨迹摘要、与功能相关的行为背景及证据标识符；{evidence_catalog}；以及{retrieved_memory}，由统一记忆库中最近确认的{k_memory}条预测—结果记录构成。每条检索到的记忆都包含较早的领域特定预后评估、随后观察到的症状状态，以及对较早评估和支持性证据与后续病程之间关系的反思。如果某条检索记录有助于当前推理，可提供可引用的既往案例标识符 [PC-. . .]。**背景/情境**。 所有{n_domains}个领域的{domain_knowledge}；每个领域的{label_cutoffs}；{ema_variable_definitions}；以及{label_set}。患者代理轨迹是在单独的前序调用中使用与 CHORAL 相同的患者代理模板生成的。**规则**。这是一项预后任务，而非仅描述当前状态。优先考虑趋势、复发、不稳定性和相对于基线的变化等纵向信号，而非孤立观察结果。评估每个领域的支持性证据和削弱性证据。应从各领域自身的角度进行评估，不要将一个领域的结论直接套用于另一个领域。将当前观察到的状态作为每个领域的默认结果；只有在引用至少一个升级或降级信号及其证据标识符时，才偏离该默认结果。根据证据标识符支撑针对患者的解释，不要编造标识符。当检索到的记忆有助于推理时，应结合较早的评估、随后观察到的症状状态及反思进行解读，并引用相应的 [PC-. . .] 标识符。为每个领域从{label_set}中确定一个标签。根据现有证据指出未解决的信息需求，但不要启动额外的患者代理检索或迭代式评估修订。不要引入独立的专科角色、编排器、智能体间讨论、专科审查或迭代修订。**输出**。 recent_behavioral_summary；domain_assessments{label, reasoning{evidence_trajectory[]， 与功能相关的行为背景, 预后展望, 症状风险推断}, supporting_evidence[]， 削弱性证据[]， 不确定性来源[],信息需求[],后续问题, 干预提示}; 跨领域关联[]；综合结果；信息缺口[]；后续问题[]；干预提示[]；引用的证据标识符[]；最终预测。

D 临床医生参与者人口统计信息

表 9。临床医生参与者的人口统计特征（$N = 22$）。临床经验包括实习、住院医师培训及后续临床实践。

参与者	Age	性别	职业身份	临床经验
P1	46	M	经认证的精神科医生	>10 年
P2	48	M	经认证的精神科医生	>10 年
P3	37	M	经认证的精神科医生	>10 年
P4	44	M	经认证的精神科医生	>10 年
P5	37	F	经认证的精神科医生	10年
P6	49	M	经认证的精神科医生	>10 年
P7	34	F	精神科住院医师	5年
P8	31	M	精神科住院医师	5年
P9	31	F	精神科住院医师	5年
P10	33	M	经认证的精神科医生	7年
P11	33	M	精神科住院医师	3年
P12	33	M	精神科住院医师	3年
P13	33	M	经认证的精神科医生	7年
P14	28	F	精神科住院医师	3年
P15	28	F	精神科住院医师	2年
P16	30	M	精神科住院医师	4年
P17	31	M	精神科住院医师	4年
P18	32	M	精神科住院医师	4年
P19	30	F	精神科住院医师	3年
P20	29	F	精神科住院医师	3年
P21	33	M	精神科住院医师	2年
P22	30	M	精神科住院医师	3年

E 临床医生问卷

表10列出了用于评估临床有用性、可操作性、工作流程适配度、可质疑性与依赖支持的问卷条目。所有条目均采用7点李克特量表进行评分。

表10。临床评估中使用的问卷项目 ation.

测量指标	问卷项目
RQ2：临床有用性、临床可操作性和工作流程适配度	
临床有用性	1. 鉴于患者近期的变化，该预后判断在临床上具有实用价值。 2. 这些支持性证据帮助我在临床上评估该预后判断的有效性。 3. 综合多个症状领域的发现，有助于解读 患者整体状态。
临床可操作性	4. 所提供的信息帮助我识别出难以通过……确定的方面 可用数据以及所需的进一步核实。 5. 所呈现的信息帮助我确定应优先处理哪些预后问题 以及原因。
工作流程适配度	6. 信息量与组织适合在实际 临床实践中进行审查。 7. 系统输出可作为 预后判断和随访规划的参照，整合到现有临床工作流程中。
研究问题***3***：可争辩性与依赖支持	
可质疑性支持	8. 我可以追溯哪些患者数据和领域特异性发现支持最终的预后 判断。 9. 我能够识别出不同领域的研究结果相互不一致，或存在替代性 解释的情况。 10. 如果我不同意系统判断，我能够识别出哪些证据或领域特定 判断应进一步接受审查或质疑。
依赖与 核实支持	11. 我认为系统的预后判断可以作为临床参考加以信赖。 12. 我能够区分系统判断中哪些部分可以依赖，哪些部分需要 进一步核实。 13. 我认为系统提供的信息不太可能导致临床上 不恰当或不安全的判断。

F 单智能体仪表板

单智能体条件使用了一个仪表板，其高层组织结构与 CHORAL 界面相似，但省略了 CHORAL 多智能体流程所特有的中间推理视图。如第 4.2.1 节所述，单智能体在一次大语言模型生成中完成了四项领域评估和综合评估，并将其与相应的支持性证据及后续信息一同呈现。图 8 展示了单智能体条件下使用的仪表板。

G 领域表现

G.1 领域级预后表现

表 11–14 报告了四个目标领域各自的平衡准确率、精确率、召回率、F1值、准确率和恶化检测。恶化指从低状态转变为高状态。总体而言，CHORAL 在抑郁（0.711）、焦虑（0.695）和压力（0.651）领域取得了最高的平衡准确率，并在抑郁和压力领域取得了最高的 F1值（0.439）。在这三个领域中，多数基线的平衡准确率仍处于机会水平（0.500），且未检测到任何恶化；持久性基线同样未检测到任何恶化，因为沿用当前标签无法预测从低状态到高状态的转变。相比之下，CHORAL 在所有

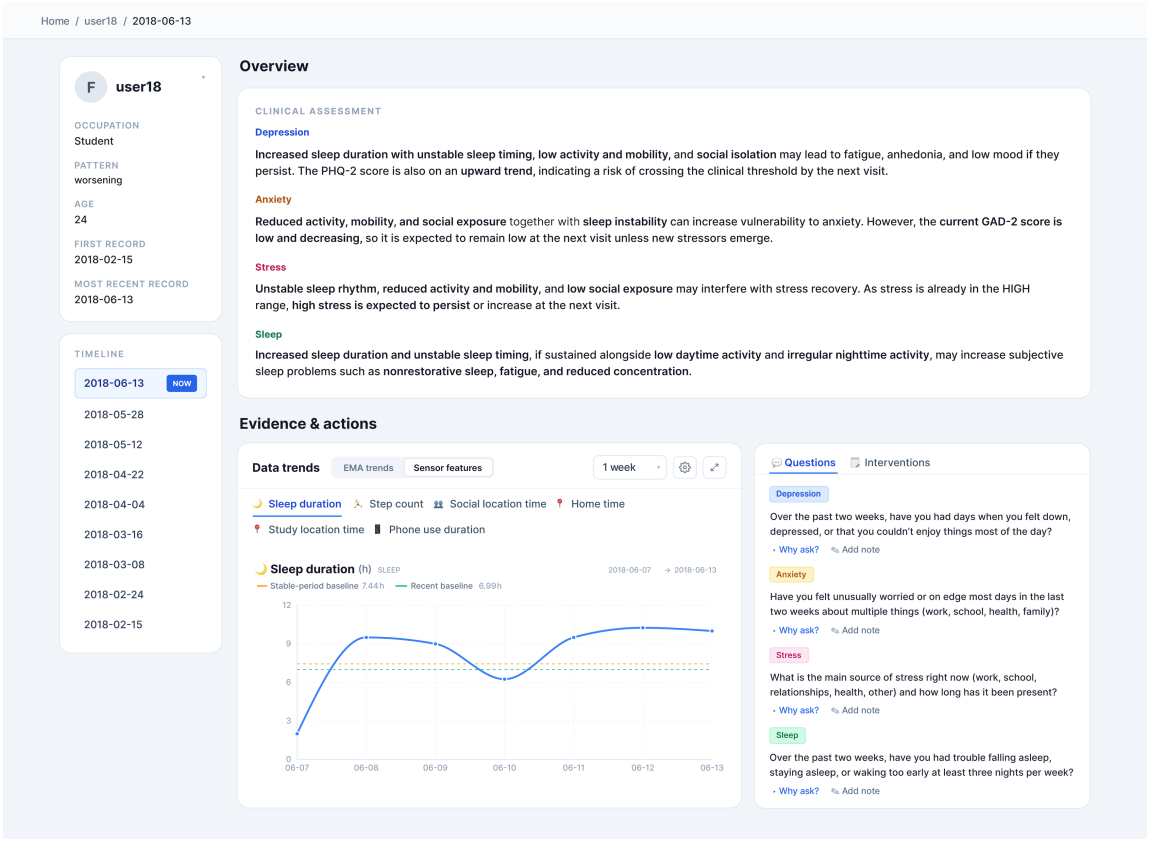


图8. 临床医生评估期间单智能体条件下使用的仪表板。

表11. 抑郁的领域级预后性能。

条件	平衡准确率	精确率	召回率	F1	准确率	恶化检测
多数基线	0.500	0.000	0.000	0.000	**0.904**	0/42 (0%)
持久性基线	0.676	**0.433**	0.408	0.420	0.892	0/42 (0%)
单智能体基线	0.689	0.295	0.507	0.373	0.836	11/42 (26%)
无编排器	0.701	0.285	**0.549**	0.375	0.824	**12/42 (29%)**
无记忆	0.681	0.352	0.451	0.395	0.868	4/42 (10%)
无按需检索	0.672	0.311	0.451	0.368	0.851	8/42 (19%)
CHORAL	**0.711**	0.387	0.507	**0.439**	0.876	9/42 (21%)

三个领域中均检测到了恶化病例。正如正文所述，睡眠是唯一一个持久性基线的平衡准确率高于 CHORAL 的领域。

恶化检测结果进一步体现了这一差异。所有基于大语言模型的条件都检测到了四个领域中的恶化病例，而多数类基线和持久性基线均未检测到任何病例。CHORAL 在抑郁、焦虑、压力和睡眠领域分别检测到 21%、21%、37% 和 58% 的恶化病例。消融结果也显示出不同的

表12。焦虑的领域级预后性能。

条件	平衡准确率	精确率	召回率	F1	准确率	恶化检测
多数基线	0.500	0.000	0.000	0.000	**0.880**	0/48 (0%)
持久性基线	0.693	**0.456**	0.461	**0.458**	0.869	0/48 (0%)
单智能体基线	0.667	0.295	0.494	0.370	0.797	7/48（15%）
无编排器	0.682	0.315	0.517	0.391	0.807	**12/48（25%）**
无记忆	0.680	0.408	0.449	0.428	0.855	1/48（2%）
无按需检索	0.678	0.296	**0.528**	0.379	0.792	**12/48 (25%)**
CHORAL	**0.695**	0.359	0.517	0.424	0.831	10/48 (21%)

表13. 压力的领域级预后性能 .

条件	平衡准确率	精确率	召回率	F1	准确率	恶化检测
多数基线	0.500	0.000	0.000	0.000	**0.792**	0/92 (0%)
持久性基线	0.621	**0.397**	0.403	0.400	0.749	0/92 (0%)
单智能体基线	0.616	0.324	0.513	0.397	0.676	26/92 (28%)
不含编排器	0.647	0.323	**0.656**	0.433	0.642	**40/92（43%）**
无记忆	0.624	0.339	0.506	0.406	0.692	18/92（20%）
无按需检索	0.606	0.306	0.526	0.387	0.653	31/92（34%）
CHORAL	**0.651**	0.341	0.617	**0.439**	0.672	34/92 (37%)

表14。睡眠领域级预后性能 .

条件	平衡准确率	精确率	召回率	F1	准确率	恶化检测
多数基线	0.500	0.000	0.000	0.000	0.576	0/113（0%）
持久性基线	**0.698**	**0.659**	0.640	**0.649**	**0.707**	0/113（0%）
单智能体基线	0.584	0.490	0.717	0.582	0.564	60/113（53%）
无编排器	0.598	0.491	0.828	0.617	0.564	73/113（65%）
无记忆	0.578	0.473	**0.879**	0.615	0.532	**80/113（71%）**
无按需检索	0.603	0.505	0.736	0.599	0.582	78/113（69%）
CHORAL	0.633	0.524	0.806	0.635	0.607	65/113（58%）

配置之间存在召回率–特异性权衡。具体而言，移除编排器通常会提高召回率，但降低特异度；而完整的 CHORAL 在四个领域中的三个领域取得了最高的平衡准确率。

H 运行间变异性分析

为评估运行间变异性，我们在相同设置下独立地重新运行了一次 CHORAL，并比较了预后性能、证据使用、领域级预测以及所引用证据标识符。证据标识符一致性基于每个病例周的证据目录中各标识符的二元引用/未引用编码进行计算（$N = 34,188$），并汇总编排器最终输出中四个领域级推理所引用的内容。

两次运行取得了相近的平衡准确率（0.673 对 0.672）和证据渠道覆盖率（89.2% 对 89.0%）。领域级预测显示出 88.7% 的一致意见（$\kappa = 0.745$），所引用证据标识符显示出 90.8% 的一致意见（$\kappa = 0.812$）。总体而言，两次运行的汇总性能相近。