arXiv:2611.29996v1 [cs.HC] 23 Nov 2026

MindBridge：用于交互式认知监测

XIAOJIE GU，东京大学，日本HAOYU ZHANG，东京大学，日本QIAN NIU，东京大学，日本FAN GAO，东京大学，日本RIO YANG，东京大学，日本
ELENI FRILIGKOU，耶鲁大学医学院，美国
KEVIN JIN，耶鲁大学医学院，美国YUTAKA MATSUO，东京大学，日本IRENE LI，东京大学，日本

我们介绍*MindBridge*，这是一款由临床医生参与设计的多语言、多模态认知监测工具，将参与者自主控制的访谈与可解释报告相结合。报告将词汇多样性和停顿模式与语音或文本证据联系起来，并披露数据质量和缺失性。在一项针对10名成年人的文字交互形成性研究中（7人年龄为18至59岁，3人年龄为60岁或以上），后续提问促使参与者进一步详述，但有时会引入需要纠正的假设。9名参与者将解释优先报告列为首选；所有参与者都认为，仅评分报告最可能诱发执着关注和过度反应。在固定的DementiaBank清单上，表现最佳的配置在有效预测上的零样本准确率达到70.3%，而整个清单上的多数类基线为56.0%。某些配置只预测一个类别；使用生成样本进行的测试揭示了不同输入类型和语言之间存在不同的错误模式。*MindBridge* 使证据收集、分析和解读过程清晰可查，并将参与者控制与监测结果的范围和局限联系起来。

CCS概念： • **以人为中心的计算 → 人机交互（人机交互）**。

其他关键词和短语：多模态界面、自适应访谈、认知监测、痴呆症研究、结果沟通、不确定性沟通、以人为中心的人工智能

ACM引用格式：
Xiaojie Gu、Haoyu Zhang、Qian Niu、Fan Gao、Rio Yang、Eleni Friligkou、Kevin Jin、Yutaka Matsuo 和 Irene Li。2018。*MindBridge*：用于交互式认知监测的多语言多模态工具。载于论文集 请务必填写版权确认邮件中的正确会议名称（会议简称'*XX*）。ACM，纽约，NY，美国，32页。https://doi.org/XXXXXXX。XXXXXXX

1 引言

认知健康会影响人们的沟通方式、日常活动管理方式，以及参与家庭和社区活动的方式生活。连贯话语，例如描述一幅图片或回顾一天的经历，能让我们一窥这些能力，因为

允许免费为个人或课堂用途制作本作品全部或部分内容的数字副本或纸质副本，前提是不得为营利或商业利益制作或分发副本，且副本首页须附上本声明及完整引文。对于本作品中由作者以外的其他方拥有版权的组成部分，必须尊重其版权。注明出处后，可以摘录。除此之外，如需复制、再版、发布到服务器或重新分发至列表，须事先获得明确许可，和／或支付费用。请向 permissions@acm.org 申请许可。© 2018 版权所有归所有者／作者所有。出版权已授权给 ACM。稿件已提交至 ACM

它需要调用词汇检索、语义知识、记忆、叙事组织、执行控制和言语产生能力 [3, 19, 25]。计算方法为研究这些能力如何体现于话语中提供了颇具前景的途径，既可以分析转录文本中的语言模式以及语音中的时间和声学模式 [6, 22]，也可以利用大语言模型（LLM）提取转录文本特征 [12]，并评估纯文本和音频文本大语言模型 [24]。这些能力为开展更便捷的多语言认知研究创造了机会，但模型性能本身并不能说明证据是如何产生的，也无法解释某项结果在合理范围内意味着什么。

现有的认知评估对话系统使用反馈语和支架式支持来促进参与和任务理解 [7, 14]。以人为中心的可解释性研究考察结果界面如何支持理解以及适当依赖 [4, 17]。因此，在认知监测中，交互在模型分析前后都很重要。除了完成任务之外，参与者还需要能够表达其意图、纠正误解，并决定要分享哪些内容。后续跟进可能有助于他们详述，但也可能要求他们在继续之前纠正某个假设，从而改变之后用于分析的回答。分析完成后，分数、不确定性和解释的呈现方式会影响读者如何理解输出 [15, 16, 18]。

挑战不仅在于解释模型的输出，还在于解释其支持性证据是如何收集和分析的。输入范围和模态特定处理会影响哪些信息能够进入摘要。将这些阶段连接起来，可以把以人为中心的可解释性延伸到监测证据的形成方式，以及该证据能够支持哪些结论。

为研究这一设计空间，我们介绍 *MindBridge*，这是一种由临床医生参与设计的多语言、多模态认知监测工具。精神科住院医师提供的结构化输入为系统的证据优先级、混杂因素披露和结果呈现提供了依据。通过有界自适应访谈，系统支持详述，并为参与者提供显式控制，以决定哪些内容进入记录。系统并不将由此产生的样本视为已经完成的输入，而是保留塑造其语音和文本证据的提示、响应及控制行动。同一证据链延伸至研究者工作台，在那里，观察结果仍与其来源模态、数据质量和缺失性相关联。总体而言，该工作流程使交互如何形成证据、多模态分析如何转换证据，以及这些证据如何支持解读变得清晰可见。

我们的评估围绕三个问题展开：

研究问题1： 有界自适应后续提问和任务级控制（重读和提示重定向）如何改变参与者对录音记录的贡献内容，以及他们对此保留了多少控制权？**研究问题2：** 当相同的样本特定证据以单纯评分、五维分解或解释优先报告的形式呈现时，人们如何解读并依赖这些证据？**研究问题3：** *MindBridge* 的多模态分析管线在固定录音和受控跨语言压力测试下表现如何？与课程、模态和语言相关的哪些局限性可能会被汇总值的表现所掩盖？

我们的评估在三个节点考察图 1 所示的工作流程。十名成年人完成了一次打字自适应访谈和两项互补的打字任务，以回答研究问题1；他们解读了两个虚构案例，并比较了三种报告格式，以回答研究问题2。针对研究问题3，我们在输入范围、模态和语言发生变化的条件下，采用 109 份英语 DementiaBank 录音 [2], 20 以及在四个预设沟通困难等级下受控的老年人角色画像，并开展中文和日语跨语言压力测试，以应用 *MindBridge* 的分析工作流程。这一设计连接了面向人类的两个边界，同时利用技术检查在分析路径中定位故障。

用户研究凸显了这样一种需求：既要支持详述，也要保留参与者纠正访谈者并控制分享内容的能力。参与者更青睐先解释证据及其局限、再呈现评分的报告。他们认为，仅评分报告最可能使人忽略评分背后的证据。稿件已提交 ACM

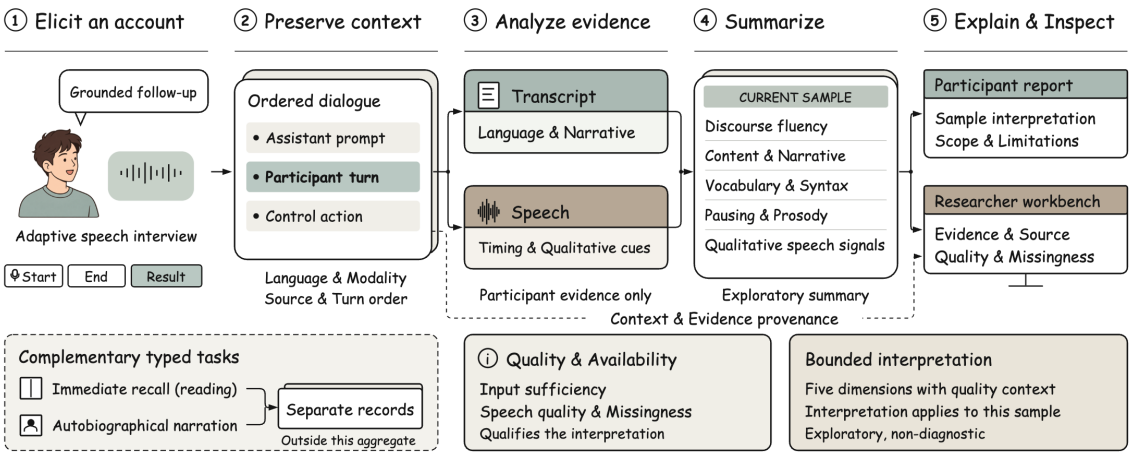


图 1. *MindBridge* 认知监测工作流程。步骤 1 和 2 通过提示引导参与者进行口述叙述，并保留提示、响应和参与者控制。步骤 3 和 4 分析转录文本和语音证据，并从五个维度概述当前样本。步骤 5 展示参与者报告和研究者工作台。两项互补的打字任务分别生成记录，用于研究问题1。

这一评分。技术检查揭示了相互补充的失败模式。在相同工作流程下，表现最佳的固定录音配置达到 70.3% 的零样本准确率，而若干配置则趋向于单一类别。精确的合成级别恢复率从文本条件下的 85.0% 降至语音条件下的 26.25%，以及语音加文本条件下的 50.0%；与此同时，中文和日语的汇总准确率相同，却掩盖了敏感度和特异度的显著差异。总体而言，这些发现表明，负责任的解读需要清晰呈现证据转换，而不只是解释最终评分。本研究有四项贡献：

• 我们介绍了 *MindBridge* 的设计与实现。这是一款由临床医生参与设计的多语言、多模态认知监测工具。它将有界自适应语音访谈和参与者控制，与面向参与者和研究者的视图相结合，展示转录文本和语音证据、数据质量及缺失性。互补的即时回忆任务和自传式叙述任务分别记录。

• 基于 *MindBridge* 及其评估，我们阐述了认知监测中以人为中心的可解释性的流程导向设计思路。解释应说明样本如何引出、哪些证据纳入分析，以及结果能够支持哪些结论，而非仅仅为最终评分提供依据。

• 我们报告了一项针对 10 名成年人的文字交互形成性研究所得的实证发现。参与者表示，他们补充了原本未计划提供的细节，并指出有些后续提问预设了他们未曾描述的事件或感受。他们更倾向于解释优先报告，但仍有人根据样本特定证据推断整体认知能力。这些发现为参与者控制和证据优先报告提供了设计启示。

• 我们通过实证研究刻画了*MindBridge*的分析工作流程，发现某些配置会将所有录音都分到同一类别；文本与语音在还原合成难度等级方面存在差异；而不同翻译和合成语言版本的准确率相同，却掩盖了各自的错误模式。在固定的 DementiaBank 清单上，表现最佳的配置在有效预测中的零样本准确率为 70.3%，相比之下，多数类基线在完整清单上的准确率为 56.0%。

2 相关工作

2.1 认知评估对话系统

言语与语言可通过词汇选择、话语组织、流畅度、韵律和发音，为认知评估提供已有的证据 [3, 10, 19, 25]。这些信号会受到信息引出情境的影响，因为图片描述、叙述、访谈和自发对话对参与者提出了不同要求 [3, 19]。因此，最终获得的样本既反映了参与者的表达，也反映了促成这种表达的交互过程。

对话代理使这种交互具有适应性。在对话式认知评估中，系统运用反馈语和支架式支持，促进参与者投入并理解任务 [7, 14]。与传统网页表单相比，对话式调查 elicited 出的回答更具体 [27]。访谈者言语也会影响自动评估结果 [21]。

语音和转录文本为多模态分析提供互补证据 [6, 22]。近期的大语言模型方法包括转录文本特征提取 [12] 和音频文本建模 [24]。Winterlight、Canary Cognitive 和 ki:elements 的 SB-C 等系统展示了基于语音的筛查与监测的实际应用 [5, 23, 26]；FRONTEO 的 KIBIT 支持专家文本分析 [11], ，相关研究则探讨了如何根据自由对话转录文本进行神经认知障碍分类 [13]。这些系统展示了语音和文本分析的实际应用。我们的重点是限制访谈者的影响、保留参与者控制权，并在样本进入语音和文本分析时保留信息引出的语境。

2.2 以人为中心的可解释性

以人为中心的可解释性关注人们能否理解人工智能支持的结果、识别其局限，并判断何时应当依赖它。相关指南强调，应清晰呈现系统的能力和不确定性，并根据用户的问题和目标来设计解释 [1, 17]。在健康相关场景中，数字和可视化格式会影响风险解读 [18], ，而不确定性展示仍取决于任务和情境 [15, 16]。仅靠解释未必能避免过度依赖 [4]。

交互式认知监测还必须考虑提示、修复、复核选项和输入模态如何构成证据，以及结果顺序如何影响对证据的解读。因此，有必要将每项结论与支撑它的交互、质量、缺失性和模态联系起来。*MindBridge* 会在信息引出、分析和结果沟通的整个过程中考察这种联系，而不是等结果生成后才补充解释。

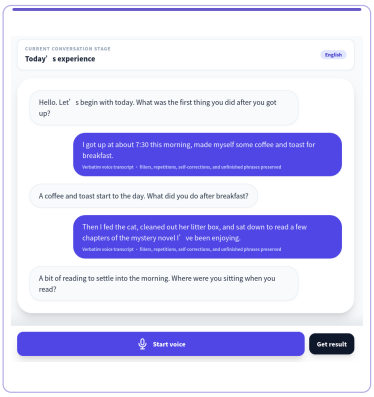
3 *MindBridge*：交互式多模态认知监测工具

MindBridge 是一款基于网页、由临床医生参与设计的多语言认知监测工具，可收集语音和文本样本，并总结可观察到的交流模式。它旨在用于认知监测研究和自我反思，而非诊断或临床评估。当前原型会对会话中提供的样本生成探索性摘要；它不会判定一个人的认知状态，也未经临床验证。

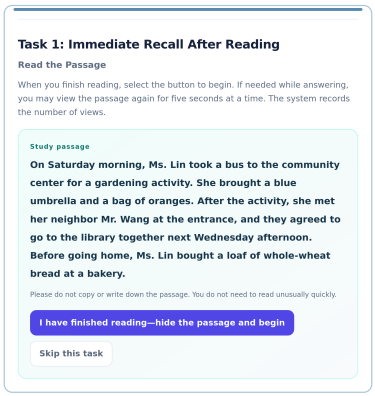
图 1 所示的工作流程包括一次受限的自适应语音访谈，以及两个相互关联的文字输入模块：阅读后的即时回忆和自传式叙述（图2(a–c)）。访谈提供用于参与者报告分析的样本。文字输入模块保留特定任务的回答和参与者控制行动，用于研究问题1，但会单独存储，不纳入当前多模态汇总。当参与者请求结果时，*MindBridge* 会检查访谈输入是否充分，分析转录文本和语音证据，并将可用的观察结果整理为五个源自经审查特征注册表的面向参与者的维度。数据 稿件 已提交至 ACM

质量和缺失性会单独报告，以限定解读范围。参与者视图说明当前样本展现了什么，以及无法确立什么。关联的研究者工作台展示有序对话、输入模态、数据质量、缺失性、特征证据和分析版本。

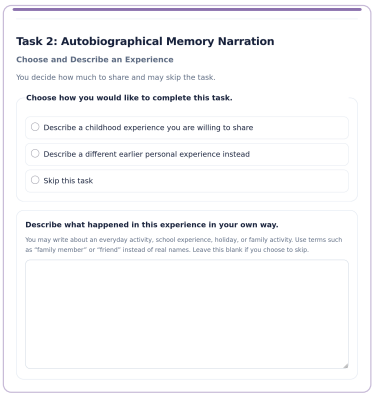
访谈和打字任务



(a) 自适应访谈

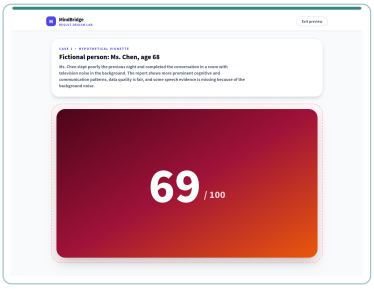


(b) 即时回忆



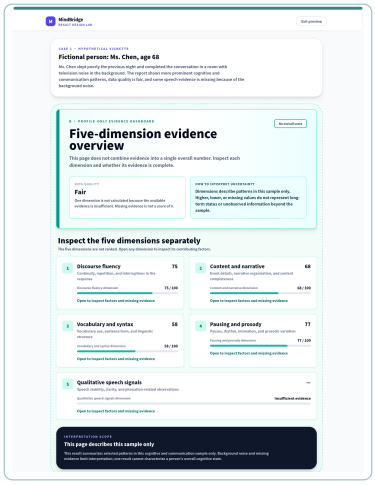
(c) 自传式叙述

案例 H 的报告方案

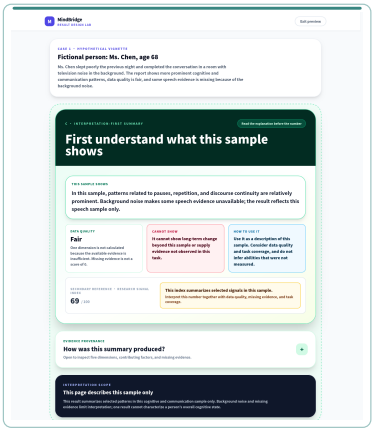


(d) 仅评分报告

MindBridge
参与者控制的认知监测器



(e) 五维报告



(f) 解释优先报告

图 2. *MindBridge*在用户研究中对比的引导式信息引出界面与报告套件。(a)：在语音原型中，参与者控制录音并回答后续提问。(b)：他们阅读一篇文章，然后复述其内容。(c)：他们选择一段个人经历进行叙述，也可以跳过该任务。(d–f) 面板在每个结果面板上方显示相同的案例 H 小 vignette。(d)：A套餐的结果面板仅显示总体指数。(e)：B套餐呈现五个维度和可展开证据。(f)：C套餐先解释样本所显示的内容及其局限，再呈现指数。

MindBridg e 通过让信息引出、分析和结果协同，支持整个监测工作流中的以人为中心的可解释性，
sis，及结果 co ups

通信过程可追溯。在信息引出过程中，**有界自适应性**指示模型以参与者提供的细节为依据，并保留每条提示，从而使偏离这一约束的情况仍可供检查。

参与者控制的轮流发言让参与者决定一次发言何时开始和结束，而**可理解的恢复机制**则展示系统状态和重试选项。在分析和结果沟通期间，**证据来源**使参与者的语音、助手提示、转录文本、语音测量和模型观察结果彼此可区分，同时**范围优先通信**将来源、质量和缺失性带入结果视图。总体而言，这些机制清晰呈现了每个样本是如何被引出、转换和汇总的。

本节其余部分将介绍监测范围与证据模型（第3.1节）、有界自适应访谈（第3.2节）、补充性记忆与叙事任务（第3.3节），以及多模态分析与结果沟通（第3.4节）。下文介绍的临床协作为监测主张的范围、证据术语、特征优先级、混杂因素警示以及参与者说明提供了依据。第5.5节中的技术检查将测试这一已实现的工作流程在不同输入模态、课程和语言下的表现。

3.1 界定监测范围与证据模型

为确定语音与语言监测工具能够负责任地报告哪些内容，我们在实施证据模型之前征求了临床意见。最初的设计咨询邀请了一位精神科住院医师（MD、MSc）参与，此人具有临床实践、神经科学和健康数据研究经验。该医生强调，语音只是痴呆评估的一个组成部分。临床解读还取决于纵向认知衰退、日常功能变化、患者及家族史、认知筛查、行为与情感观察，以及年龄、教育程度、语言、听力、情绪、药物、疲劳和共病等其他可能的解释。*MindBridge*因此将语音与语言模式视为针对具体样本的监测证据，而非独立的认知评估。这一范围也要求界面清晰呈现潜在混杂因素和证据局限。

随后，我们为监测证据模型编制了包含 11 个言语与语言特征类别的清单，并将其发送给医生，以便其异步进行结构化标注。材料介绍了二元语料库分类，说明缺少患者病史，并附有带标签的言语示例。听取录音后，医生针对每个特征类别作出三项结构化判断，共 33 项判断，涵盖临床检查相关性、与 MoCA/MMSE 的直接对应情况 [9, 20], 以及预期计算产出。医生为每项“可能”或“否”判断提供了书面理由，并在考量清单和录音后，指出没有明显遗漏的其他特征类别（表1）。实施完成后，医生审阅了包含面向参与者的维度和摘要的原型用户界面。因此，临床参与涵盖了问题界定、特征优先级排序、语音样本复核和界面审阅。这种合作为监测设计提供了指导，但并不构成临床验证。

医生提供的序数判断确立了特征注册表中的优先级。“是”表示该类别可纳入核心证据，“可能”表示仅限作为支持性证据，而“否”则将其排除在面向参与者的汇总值之外。注册表还保留了相应的理由和混杂因素警示。为落实这一排序，我们为四个“是”类别设定固定启发式权重 1，为六个“可能”类别设定权重 0.5。类别优先级及其理由由医生提供；将其转换为数值则是我们的设计选择。我们将保留的十个类别归入五个面向参与者的维度：话语流畅度、内容与叙事、词汇与句法、停顿与韵律，以及定性言语信号。随后，医生审阅了这些内容在原型界面中的呈现。

这些标注提出了监测设计的三项要求。首先，对于人物难以评估但易于计算的指标，必须保留其来源标注，而不能因自动化而被赋予临床权威。稿件已提交至 ACM

表1。一名精神科住院医师对11类言语和语言特征类别进行的结构化标注。“检查”和“筛查”再现了该医生对临床检查相关性以及对MoCA/MMSE直接覆盖范围的判断。“产出”表示预期的分析价值。最后一列记录了我们随后进行的设计转化。

特征类别	Exam	筛查	产出	复习依据与设计转化
词汇多样性	No	No	可能	人工评估较为困难，但通过计算可能具有信息价值；支持性登记特征。
句法复杂度	No	No	也许	人工评估较困难，但通过计算分析可能具有参考价值；支持性登记特征。
话语流畅度	Yes	Yes	Yes	通常在临床上具有重要意义；核心登记特征。
停顿与言语节律	Yes	No	可能	具有临床参考价值，但可能受到抑郁症的混杂影响；作为辅助特征，并披露这一混杂因素。
韵律	Yes	No	可能	具有临床参考价值，但可能受到抑郁症或精神病性障碍的混杂影响；作为辅助特征，并披露其混杂因素。
一般声学特征	No	No	No	可能受到年龄、种族和社会经济地位的混杂影响；未纳入面向参与者的汇总。
言语不流畅	Yes	Yes	Yes	在常规临床实践中具有重要意义；是核心登记特征。
语言结构	No	No	也许	语速可在临床上观察到，而其他结构则难以手动量化；抑郁症可能会影响语速；属于辅助特征。
内容质量	Yes	No	Yes	叙述在临床上的显著完整性和相关性；核心登记库特征。
高层次文本特征	Yes	No	Yes	叙述在临床上的显著全局特征；核心登记特征。
临床声学生物标志物	No	No	可能	难以人工评估，但在计算分析中可能具有信息价值；保留作为辅助性定性语音证据，并呈现为定性言语信号；不声称其为临床生物标志物。

其次，易受精神疾病或人口统计学混杂影响的特征必须明确标注局限性，且不应在汇总值中占据主导地位。第三，未纳入 MoCA 或 MMSE 并不意味着某项特征缺乏研究价值；相反，界面应区分常规临床观察、计算证据和不确定性。在实现中，不可用的证据仍记为 null，并从分母中剔除，而不是视为观测值为零。模型派生观察为特征层面提供证据，而确定性规则则根据带版本信息的登记库计算各维度及总体汇总值。这种区分使我们能够辨别临床优先级设定、模型观察结果和确定性汇总各自对结果的贡献。

3.2 通过有界自适应访谈收集参与者特定叙述

自适应访谈会收集参与者关于会话当天活动的语音叙述。虽然所有参与者都会收到相同的开场问题，但后续提问会根据每位参与者提及的事件和细节进行调整。进入时，参与者输入自己的编号，选择中文、英语或日语，并复习在线知情同意页面。只有在参与者选择同意后，系统才会创建会话。系统会以所选语言提出开场问题，询问参与者当天起床后的第一项活动。这一主题可以在熟悉的叙事中引出时间、空间、社交和事件细节，而无需引入一系列彼此无关的认知任务。

访谈通过一个由参与者控制、反复进行的循环展开（表2）。参与者选择麦克风控件开始一个轮次，再次选择该控件以提交轮次。在此过程中，界面会持续显示状态反馈，包括连接中、可以开始说话、正在接收语音、等待回复以及失败状态。参与者完成的转录文本和助手回复会以角色区分的聊天气泡形式呈现，让参与者能够查看对话历史以及叙事的当前进展。当前原型支持参与者语音输入，而非完全的语音到语音交互：参与者通过语音发言，助手的回答则以文本形式显示。图2(a)展示了参与者完成两个示例语音轮次后的界面状态。固定的开场白、标注清晰的语音转录文本以及每次仅包含一个问题的后续提问，都会作为交互背景持续显示；持久控件则让参与者决定何时开始下一轮或请求结果。

表2。参与者控制的交互循环及所保留的记录由 *MindBridge*。

阶段	参与者操作	系统响应	保留记录
进入	输入其编号，选择语言，复习同意书页面，并选择同意	创建会话并显示一个固定的开放式问题	参与者提供的 ID、语言和同意记录
说话	开始 the 麦克风，发言并结束轮次手动	显示当前捕获状态并提交一个有界语音轮次	有序语音片段和明确的轮次边界
后续跟进	阅读回复并决定接下来描述什么	应用一项提示策略，要求提出一个基于情境的问题，且该问题基于参与者的——参与者的叙述，并异步显示文字记录	参与者文字记录、助手回复、角色、顺序和时间
修复	重复一个轮次，或在……后重试一个可见故障	使用中性重复请求，并在可见重试期间保留该贡献	修复请求、重复贡-献和轮次顺序
完成	请求结果或继续说话	检查数据充分性并区分保存与分析	数据质量状态、证据来源信息和报告版本

MindBridge 采用由参与者控制的轮次边界：只有在参与者明确开始一个轮次后才会采集语音，且只有在参与者结束该轮次时才会提交。访谈者会结合原始语音和先前对话作出回答，而独立的转录分支则保留文本记录。后续的监测分析可以利用语音和转录文本证据，同时避免混淆二者的来源。这种架构保留了即时的对话循环和可追溯的证据记录。部署的实现使用 gpt-realtime-2 进行语音交互并生成定性语音观察。GPT-5.4 生成文字后续提问、分析存储的转录文本，并执行最终的结构化证据融合步骤。它会分配特征级分数，随后由服务器以确定性方式计算五个维度分数和总体研究信号指数。技术评估沿用了这一分析工作流程，同时改变模型配置、输入范围、模态和语言。

由于后续跟进问题可能改变之后用于分析的样本，自适应策略有意保持狭窄。访谈者始终围绕参与者从醒来到现在的经历展开，并可跟进参与者提及的具体锚点，例如行动、地点、人物、物体、旅程、事件顺序或明确表达的感受。每次正常回复只采用一种主要对话动作，通常由一两句简短的话组成，并提出一个负担较轻的问题。该策略禁止推断未明说的日常习惯或情绪、称赞参与者的表现、连续提出多个问题，或引入图片描述、听力测试及其他评估任务。如果参与者记不起来或拒绝回答，访谈者会毫无压力地转到较晚的时间点。如果语音无法辨识，访谈者会中性地请求对方重说，而不是猜测内容或声学原因。这些提示约束旨在让自适应过程回应参与者，同时限制智能体对信息引出任务的改变程度。

参与者可随时要求完成。生成监测结果之前，系统会检查会话是否包含足够的口语材料和可用的非静音语音。此项检查可避免将材料稀少或无法使用的录音呈现为可解读的结果。如果未满足要求，界面会说明缺少哪些材料，并允许参与者继续。条件满足后，应用程序会提供确定性收尾，同时仍允许参与者补充内容。在完成过程中，界面会区分保存与分析，并在发生故障后提供重试或安全退出选项。

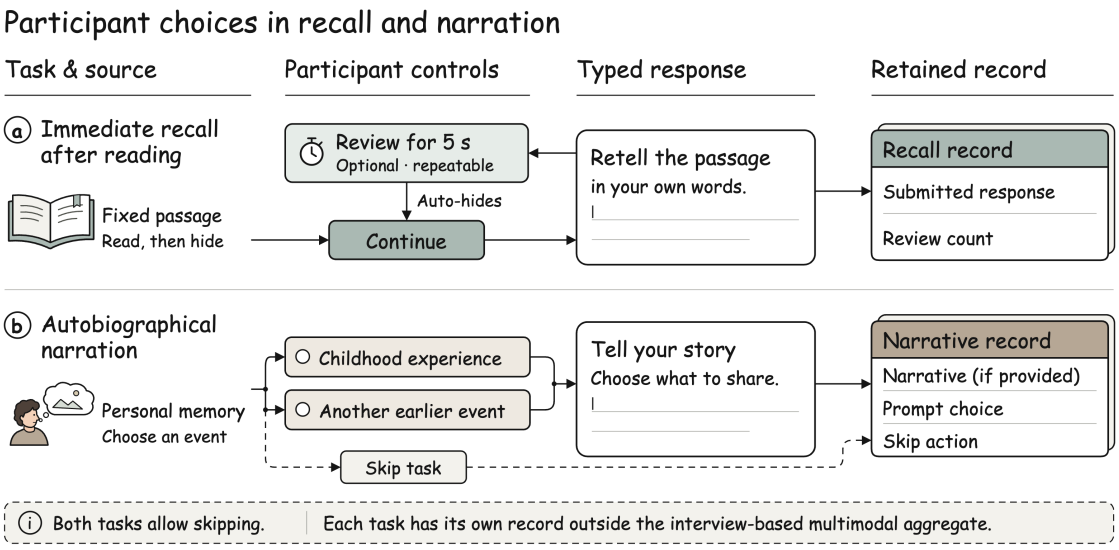


图 3. *MindBridge* 中的两项互补打字任务，展示了每项任务的来源、参与者控件、回答和保留记录。(a)：阅读后进行即时回忆，要求重构一篇固定短文，并可选择进行五秒复习；系统会保留提交的回答和复习次数。(b)：自传式叙述允许参与者选择童年或其他较早的经历，并决定分享哪些内容；系统会保留提示选择、跳过行动，以及参与者提供的叙事。两项任务均允许跳过，并在基于访谈的多模态汇总之外生成单独的记录。

3.3 收集互补的记忆与叙事样本

单一的对话形式只能展现参与者表达的一个侧面。我们选择了两项互补任务：阅读后的即时回忆和自传式记忆叙述。医生随后审阅了这些任务旨在引出的记忆与语言特征的临床相关性，并认为它们是自适应访谈的合适补充。这些任务在记忆来源和参与者可采用的控制方式方面有所不同（图 3）。它们是独立的文字输入活动，其任务专属记录可用于研究问题1。这些记录与语音输入访谈分开存储，也不会纳入当前面向参与者的汇总。

阅读后的即时回忆会呈现一篇简短文章，并要求参与者重构其内容；是否使用可选的五秒复习会与回答一并记录，而不会被视为错误。自传体记忆叙述会引导参与者回忆童年或其他更早期的经历，同时由参与者选择要披露的细节，也可以跳过该任务。这些机制会将可选的复习选择作为来源信息予以保留，并在回答环节明确披露和跳过控件，而不是将每项任务呈现为静态回答框。与自适应访谈结合后，这些任务为研究工作流程提供了三条互补的信息引出途径，涵盖近期呈现的信息、以个人经历为基础的回忆，以及逐步展开的日常叙述。在图2中，面板(b)展示了任务1的文章和阅读控件，面板(c)展示了任务2的提示选择。附录A在图6和图7中提供了更大的任务视图。

3.4 分析多模态证据并传达监测结果

参与者请求结果且满足质量要求后，*MindBridge* 会从三个来源构建会话级证据记录：确定性的时间性言语测量指标、对言语信号的定性观察，以及从参与者转录文本中提取的语言和叙事证据。助理轮次会作为信息引出上下文保留，但不会被视为参与者证据。可用观察结果会映射到上述医生标注的特征注册表，并组织为五个面向参与者的维度：话语流畅度、内容与叙事、词汇与句法、停顿与韵律，以及定性言语信号。特征的适用性、优先级、依据和混杂因素披露均会保留在其来源中。不可用的证据会保持缺失状态，而不会被转换为零，并会从汇总中排除。录音质量检查会确定证据的可用性并限定信心；这些检查会单独报告，而不会被视为额外的认知证据来源。

参与者结果页面呈现当前样本的探索性认知监测摘要，并说明其质量局限。每项保留特征都会获得一个针对该样本的 0 到 100 分，分数越高表示阿尔茨海默病相关特征清单中以困难为导向的模式越显著，而非认知表现越好、疾病概率越高或诊断结果。*MindBridge* 将各维度分数和总体研究信号指数计算为现有特征分数的注册表优先级加权均值。核心特征的权重为 1，辅助特征的权重为 0.5。所得指数不代表一般认知能力、诊断结果或痴呆症概率。页面同时呈现此范围、数据质量、缺失证据和五个维度，以确保结果始终与会话中提供的材料相关。附录 C 详细说明聚合规则。

为考察呈现方式如何影响对相同监测证据的解读，我们为用户研究设计了三种报告方案（图 2(d–f)）。三种方案均在结果面板上方显示相同的案例简介。在仅显示分数的基线方案中，面板仅显示总体 0–100 指数。五维设计移除了汇总指数，并要求读者综合维度和质量级别方面的证据。解释优先设计保留该指数作为次要参考，但会先说明样本显示了什么、哪些内容仍未知，以及质量和缺失证据如何限制解读，再呈现数字。仅显示分数的方案有意将呈现内容限定为信息最少的数值，而另外两种方案则提供更丰富的解释形式。完整的研究材料详见第 4 节。

这些面向参与者的替代方案，与一个面向研究人员的工作台相辅相成。该工作台展示每个会话的有序对话、输入模态、数据质量、缺失性、特征级证据和组件版本。研究人员可以检查每条观测的来源模态，并复查同一会话中的有序对话。面向研究人员的报告视图和有序对话视图见附录A（图 8 和图 9）。

4 信息引出与解读的形成性研究

十位参与者完成了一项包含两个部分的在线研究。首先，他们使用*MindBridge*的文字交互完成了一段限定式对话、阅读后的即时回忆任务，以及自传式叙述（研究问题1）。其次，他们解读了两个虚构案例的报告，对三套报告方案进行评分和比较，并完成书面总结（研究问题2）。这些虚构报告是固定的研究材料，并非根据参与者自己的回答生成。表 3 展示了研究流程以及每个步骤收集的数据。

表 3. 研究流程及各步骤收集的数据。案例 H 和案例 L 分别是样本级信号指数相对较高和较低的虚构案例。A套餐、B套餐和C套餐分别提供仅分数、五维度和解释优先报告。

Step	参与者活动	收集的数据
1. 入门	复习了单样本范围，收到报告套餐分配，并报告了年龄段和技术熟悉度	参与者提供的 ID、分配的套餐、年龄段、阅读语言、人工智能使用频率和网络信心
2. 有界聊天	输入了当天的叙述，并回答了后续问题，完成了聊天问卷	有序对话、14 项条目级评分、一个有帮助的后续示例、报告了互动问题，并报告了对预期措辞所做的修改
3. 互补任务	重构最近阅读的文章，并进行可选复习，然后讲述一段童年或其他更早的经历	复习次数、自由回忆与线索回忆、提示选择、叙述—体验字段、八项任务体验评分，以及开放式反馈 back
4. 第一份报告：案例 H	评定报告前任务重要性，查看了指定报告，并用自己的词语进行解释，并完成后续事项	含义、局限、使用理由、质量回忆、重要性变化、信心和影响因素
5. 第二份报告：案例 L	完成报告解读任务后再次进行完成案例 H 的整个部分	与案例 H 相同的测量项目，作为第二案例处理响应
6. 界面评价	完成两个案例后，对报告界面进行评价	六个清晰度/可访问性条目和五个特定能力依赖条目
7. 套餐比较-son	比较了案例 H 的全部三种套餐，并根据其支持正确理解的程度对它们进行排名	对可解释性、分数固着、清晰了解不确定性、担忧或过度反应，以及参与者偏好的判断报告和首屏要求
8. 事后总结	回顾了聊天与报告交互	围绕首次关注、报告解释，提交了五条文字回答，内容涉及质量与不确定性、后续体验以及期望的修订

4.1 研究设计、参与者和范围

有意采用文字输入，能够排除麦克风、轮次结束、自动语音识别和声学变化对信息引出研究的影响，从而使研究问题1能够聚焦于后续提问相关性和任务层面的自主性。研究问题2则单独考察参与者对完整信息包的报告理解。在形成性研究版本中，GPT-5.4会生成有界追问，参与者可以在发送前修改，因此信息引出研究的结果反映的是与智能体进行文字交互的情况。本研究的范围因此限定于文字交互配置，所用固定虚构报告并非根据参与者的对话生成。语音路径的评估留待未来工作。

我们通过社交媒体招募了中文或日语母语者。参与者在线完成研究，并按照略高于适用的当地法定最低工资的实际时薪获得报酬。

我们分析了全部 10 份已完成的参与者记录，参与者年龄为 18–80+ ，对话式人工智能使用经验各不相同。所有参与者均选择中文作为舒适的阅读语言；其中四人还选择了英语，一人选择了日语。表 4 报告了参与者层面的年龄、技术使用经验和报告分配情况。我们按照会话创建顺序分配 P1–P10；这些论文编号不会暴露参与者提供的标识符。

表 4。参与者背景与完成的文本聊天记录。人工智能使用为自我报告频率；网络信心按 1–5 分评定。“单元”是原型的与语言无关的类词单元计数，并非回答质量的衡量指标。“第一份报告”表示为第一个虚构案例分配的套餐，之后才对全部三个套餐进行比较。A套餐仅提供分数，B套餐提供五维度分解，C套餐则在指数之前提供解读。

ID	年龄段	第一份报告	对话式人工智能使用	网络信心	用户轮次	单元
P1	18–29	B套餐	几乎每天	5	6	97
P2	18–29	C套餐	每周数次	4	6	141
P3	30–44	B套餐	每月数次	4	6	152
P4	45–59	B套餐	少于每月一次	3	6	122
P5	45–59	C套餐	从不	2	6	142
P6	60–69	C套餐	每月数次	4	6	203
P7	70–79	A套餐	从不	2	6	79
P8	30–44	A套餐	几乎每天	5	7	60
P9	30–44	B套餐	几乎每天	5	6	102
P10	80+	B套餐	不到每月一次	3	6	128

参与者仅在完成在线知情同意后方可进入，提供一个 ID 以关联研究模块；对于原本会在研究团队专用服务器上保留六个月的记录，参与者可以申请提前删除。问卷将所有报告标注为虚构内容，将*MindBridge* 描述为研究原型，并指出单个样本无法反映整体认知状态。下文引文使用 P标识符。

4.2 引导边界：后续提问与任务级控制

为回答研究问题1，我们采用有界聊天和两项互补任务，考察参与者贡献内容的变化及其对回答的控制。

研究专用对话体现了第3节中的有限策略。它要求模型每次针对参与者对当天的叙述中的某个细节提出一个后续问题。参与者可以修改回答、选择披露哪些内容、跳过或结束。角色分离轮次和可见的系统状态保留了交互顺序。附录A重现了用于将此交互与语音采集隔离的文本输入界面（图5）。

我们评估了这些后续提问带来的引出收益和交互成本。14个研究专用五点评分项目涵盖可读性、延迟、相关性、单个问题的清晰度、补充细节、控制、修复、评价顾虑、努力程度和重复意愿。一个开放式条目用于收集有帮助的后续提问示例，另有两份检查清单用于记录互动问题，以及参与者是否删除、缩短、重组、添加或改写内容。这些是形成性条目，并非经验证的量表，也不用于与非自适应访谈者进行比较。

相同的参与者完成了第3节所述的两项互补打字任务。即时回忆会在回答前隐藏一篇固定短文，并记录可选的五秒复习，以及重构、细节回答、评分和反馈。自传式叙述会记录参与者是否使用童年提示、对其进行转向，或跳过提示，同时记录其叙事字段、评分和反馈。我们将这两个模块作为探测任务交互和证据来源的工具进行分析，而非用于衡量记忆准确性或个体层面的能力。

表 5. 报告解读研究中使用的两个虚构案例。案例 H 和案例 L 表示样本级信号指数相对较高和较低，而非整体认知能力的级别。

Case	案例背景	解读问题
案例 H	信号指数相对较高；睡眠状况不佳、电视噪声，数据质量一般，且缺失语音证据	读者是否会将高指数视为定论尽管数据质量有限且缺失证据。
案例 L	信号指数相对较低；数据质量良好，但仅是熟悉日常活动的叙述，没有独立记忆，注意力或推理任务	无论读者是否据此推断出良好的整体认知功能-仍有待考量。

4.3 解读边界：报告层级结构

参与者查看了两个虚构案例，其样本级信号指数分别较高（案例 H）和较低（案例 L）。每个案例均通过三种报告方案之一呈现：A套餐、B套餐或 C套餐。表 5 总结了案例设置及其旨在探测测试的解读问题。

醒目的数字可能会在读者查看其来源、质量或局限性之前影响解读。因此，我们在每个案例中，基于相同证据构建了三种完整的呈现套餐并进行比较。在共同情境材料下方，A套餐仅显示汇总值。B套餐突出显示质量、维度和可展开证据。C套餐则将样本范围和局限性置于次级指数之前。由于信息、层级、汇总、披露和样式同时发生变化，因此无法仅将响应归因于评分位置。图 2(d–f) 并排展示了案例 H 的三种默认界面。

每个案例首先要求参与者向一位虚构的家庭成员作出开放式说明，并将其编码为限定于样本范围，或将该数值转化为总体认知能力或临床结论。后续条目考察数值含义、对质量或不确定性的回忆、与证据一致的使用、清晰度、可访问性以及特定能力依赖。在报告呈现前和完整案例模块结束后，我们分别测量了感知到的证据重要性，因此其变化不能归因于报告页面本身。随后，参与者比较并排列了A套餐、B套餐和C套餐的顺序，并完成开放式总结；我们没有将这些结果合并为通用的信任或可用性评分。

4.4 研究流程与分析

在报告解读阶段，研究工具为案例 H中的两名、五名和三名参与者分别分配了A套餐、B套餐和C套餐。案例 L中有九名参与者保留了最初的套餐；后续的研究工具配置为案例 H中的P9分配B套餐，为案例 L分配C套餐，从而使案例 L中的参与者人数分别为两名、四名和四名。这些非区组分配以及配置差异带来了多样化的接触，而非界面效果比较。落地页首先说明了单样本范围，以及质量、缺失性和情境的作用。对于每个案例，参与者先填写开放式说明，随后回答质量回忆题及其他封闭式题项。案例 L按照固定顺序排在案例 H之后，并被视为第二案例回答。随后，参与者对界面功能进行评价，比较全部三个套餐，并完成文字总结。这种比较支持形成性评议，而非参与者内因果检验；研究未另外收集口头访谈或引导员笔记语料。

系统不记录停留时间、披露内容展开、滚动或查看顺序；首次关注属于自我报告，而非通过注视或日志测量获得。

在分析中，我们将五点评分项目报告为中位数 [范围], , 将分类响应报告为 n/N，并将感知显著性变化计算为每位参与者的后测值减去前测值。对于互补任务，我们报告复习和提示选择操作、相同的项目级汇总，以及开放式反馈中明确反复出现的顾虑。我们未计算回忆

评分，未按年龄比较任务表现，也未将回答内容解读为认知状态。所有十条纳入记录均计入每个分母；未对无响应进行插补。鉴于首次接触组人数较少且分布不均，且采用了整体方案操控，我们未进行显著性检验、计算置信区间、比较年龄，也未提出界面优越性主张。

我们对两项开放式报告解释采用了透明的先验描述性规则。如果回答明确将页面限定于所呈现的样本或任务，或拒绝对稳定、长期或整体认知能力作出推断，则标记为样本范围内。如果回答将所显示的数值视为对虚构人物认知水平的测量，或视为临床结论，则标记为过度泛化。当回答提及数据质量、缺失证据或未测量的任务时，计为质量回忆。封闭式证据一致性使用题项单独报告，以免将选择提示答案等同于生成校准后的开放式解释。这些简单编码未经过独立的双重编码；因此，我们提供计数、参与者标识符和对比摘录，而非信度统计量。

对于聊天记录、有帮助的后续提问回应、任务反馈、检查清单和事后总结，我们采用描述性内容分析来识别交互机制和设计要求。我们根据存储的对话核查每一项所报告的无依据的假设，并保留反例。语料库仅包含参与者输入的中文文本；下文的简短摘录均为英文翻译。我们不声称已达到主题饱和，引用内容旨在具体呈现某一机制，而非暗示其普遍性。

因此，本研究支持的是机制层面的主张，而非因果层面或总体层面的主张。案例 L 是第二个案例的回答，九位参与者实现了同层级迁移；随后对 A套餐、B套餐和C套餐的比较，则支持对完整套餐进行批评，而非孤立地考察界面效应。完整的局限性见附录 D。

5 结果

在本节中，我们通过考察聊天记录、问卷回应，以及段落查看记录和提示选择记录来回答研究问题1。针对研究问题2，我们分析报告解释、参与者对依赖报告意愿的评分，以及对三种报告方案的比较与排序。针对研究问题3，我们通过固定录音和受控合成检验来考察*MindBridge*的分析流程。

参与者表示，他们会根据后续提问补充细节，并指出了引入无依据假设的问题。大多数参与者将报告解读限定在当前样本范围内，而对比案例则显示，过度概括在不同报告方案中仍普遍存在。软件包比较表明，应先解释范围和局限性，再呈现汇总分数。图 4 总结了部分问卷回应和任务操作计数。

5.1 研究问题1：后续提问与任务控制塑造记录

结构化响应说明了为何“相关”与“中性”是不同的交互属性。所有参与者都给出了最高相关性评分，九位参与者报告补充了最初未计划提及的细节；而 P9 则缩短了原本打算给出的回答，同时仍认为餐食细节提示有所帮助。有效的提示会选取参与者提出的锚点，并要求提供一个级别的具体说明。例如，P1 报告称，一个图表相关问题引出了实验、图表形式、颜色和坐标轴标签等信息。

三位参与者报告称，后续跟进中出现了无依据的假设。P2 在对话过程中明确纠正了访谈者。当被问及他们接下来是修改方法还是措辞时，P2 解释说，他们当时是在为一门课程做准备。P1 和 P6 则继续讲述，并在事后总结中指出了这一假设。P1 指出，“返回”一词暗示他们曾离开实验室。P6 指出，“意外的喜悦”赋予了他们一种从未表达过的感受。

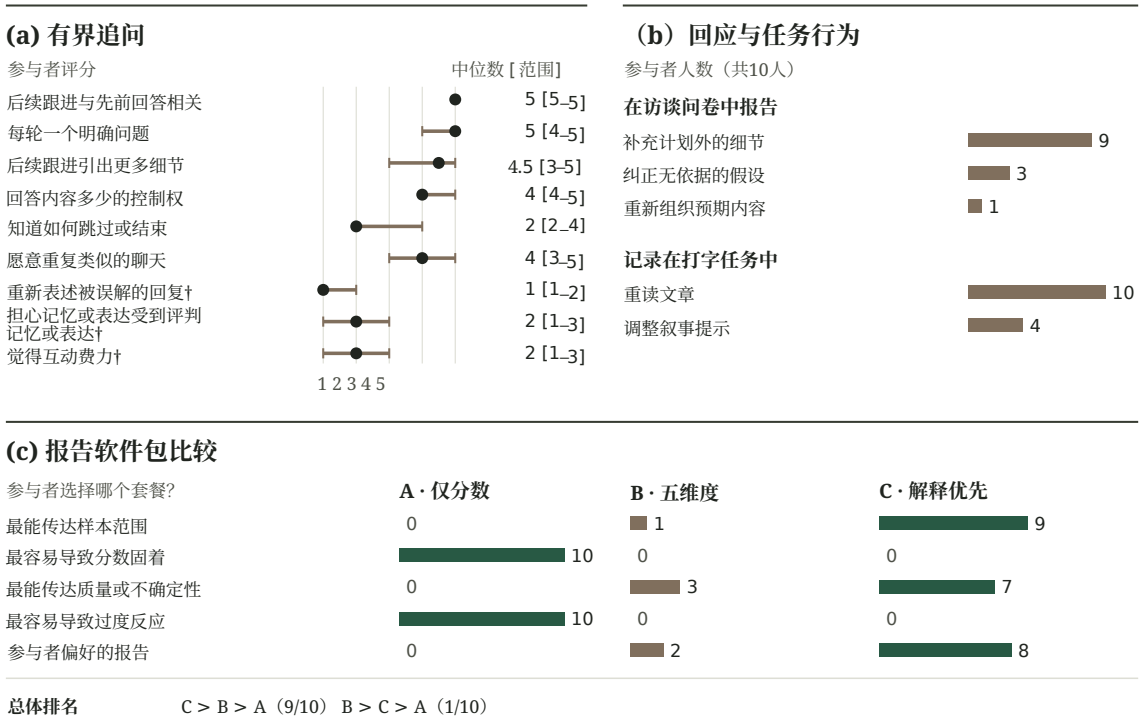


图 4. 形成性研究在信息引出和解读环节的描述性结果（N = 10）。(a)：针对有限范围内后续交互的研究专用五点评分项目的回答。圆点表示中位数，横线表示观测范围，† 标记措辞为负向的交互成本条目，分数越高表示体验越不理想。(b)：访谈问卷统计了参与者在后续提问中注意到的预期回答和假设变化。任务日志记录了段落查看记录和提示选择。(c)：参与者对三种完整报告方案进行暴露后比较的计数：A套餐（仅分数）、B套餐（五维度）和 C 套餐（解释优先）。绿色条形表示每个问题中最常被选中的套餐。最后一行按参与者认为各套餐支持正确理解的程度对其进行排名。(b) 和 (c) 面板报告的计数均以 10 人为基数。这些描述性发现刻画了本次形成性样本，但并未估计界面因果效应，也未证实某种套餐优于其他套餐。

不同界面级别上的感知能动性也不尽相同。参与者一致认为，他们能够控制回答的长度，很少报告自己需要重写，并且普遍认为任务所需努力程度较低。然而，对于是否知道如何跳过问题或结束任务，其中位数仅为 2。交互策略中虽有对照组，但在界面中并不够容易发现。总体而言，这些结果表明，应采用有依据的开放式续问，并持续提供跳过与结束功能可供性，而不是依赖参与者在对话中表达拒绝。

互补任务以不同的粒度展现了能动性。十位参与者在即时回忆中都使用了可选复习，中位复习次数为 2（范围 1–4）。任务清晰度为 4.5 [4–5], ；表达所记内容的支持度为 4[3–5], ；复核控制清晰度为 4 [2–5], ；努力程度为 3 [2–5]。五位参与者认为固定的五秒时间窗口太短或过于仓促。因此，保留复习次数有助于区分在不同刺激暴露条件下产生的响应。

十位参与者均完成了自传式叙述：六位使用了童年提示，四位对其进行了调整，没有人跳过。表达在 4 [3–5], 之间；对披露的感知控制和跳过的情况在 5 [4–5], 之间；努力程度在 3 [2–5], 之间；不适感

表6. 报告解读与特定能力依赖（$N = 10$）。案例 H 是首先展示的高指数案例；案例 L 是在案例 H 的完整区块之后展示的低指数案例。五点评分项目和后测减前测的显著性变化为中位数 [范围]。

结果	观测值
样本范围内的开放式解读，案例 H	8/10
过度概括为个体层面的认知能力，案例 H	2/10
回忆质量、缺失性或不确定性，案例 H	9/10
选择与证据一致的用法 , 案例 H	10/10
感知显著性变化，案例 H	+1 [0–2]
解读信心，案例 H	4 [3–5]
样本范围内的开放式解读，案例 L（第二个案例）	8/10
过度概括为个体层面的认知能力，案例 L	2/10
回忆质量、缺失情况或未测量任务，案例 L	8/10
选择与证据一致的用法，案例 L（第二个案例）	8/10
感知显著性变化，案例 L	+1 [0–2]
解读信心，案例 L	4.5 [3–5]
依赖单个样本总结的意愿	4 [2–5]
依赖标记缺失或低质量数据的意愿	5 [1–5]
依赖比较重复样本的意愿	4 [2–5]
依赖推断页面未提供信息的意愿	1 [1–5]
依赖无证据支持的总体评分的意愿	1 [1–5]

为 1 [1–2]。任务要求讲述故事、其背景以及一个令人难忘的细节。七位参与者报告称，这些部分之间存在重叠。

事后总结揭示了后续提问中存在的问题，即使参与者在未质疑问题的情况下作答也是如此。参与者通过回看文本和提示重定向来控制其任务回答，尽管其中一些人不确定如何跳过或结束访谈。

5.2 研究问题2：报告理解与特定能力依赖

表6总结了参与者如何解读每个案例，以及他们愿意依赖哪些报告能力。共同情境材料还提供了有关数据质量或任务范围的线索，参与者可以在回答回忆问题时加以利用。图4(c)展示了他们在比较案例 H 的全部三个报告套件后的判断。

八名参与者在高指数案例（案例 H）和低指数案例（案例 L）中，都将其解读限定在当前样本范围内。最初被分配到仅评分报告（A套餐）的两名参与者形成了有益的对比。P7谈到A套餐时说：“我只能看到69，不知道这个评分由什么构成”，随后拒绝将其解读扩展到这个有噪声的样本之外。相比之下，P8将69描述为一个“及格但不算特别好”的认知水平，后来又将23解读为需要就医的认知衰退。P8表示，在不核查证据的情况下，其依赖总体分数的意愿最高，并且对案例 L 的解释非常有信心。P7表示，其依赖该分数的意愿最低，信心也较低。因此，单独的数字并不会决定误解；它会造成一个解释真空，使既有的分数惯例得以占据主导。

P9 在更丰富的层级结构下提供了一个互补的反例。在查看高指数案例（案例 H）的五维报告（B套餐）后，P9 承认存在噪声和缺失的语音证据，但仍将该页面描述为人物的“认知测试水平”。在查看低指数案例（案例 L）的解释优先报告（C套餐）后，尽管前文已说明范围和缺失任务，P9 仍将低值解读为需要医疗关注的认知问题。P9 还报告称，其最愿意推断缺失信息，并依赖整体稿件提交至 ACM

评分，而不考虑其证据。因此，解释顺序可以支持有边界的解读，但无法保证这一点；先前的预期仍然是解读边界的一部分。

测量结果的排列揭示了第二个区别。P8 和 P9 在针对案例 H 给出人物层面的开放式解读后，都选择了与证据一致的封闭式答案，因此提示后的案例 H 结果为 10/10，尽管只有八份开放式解释以样本为范围。两人都为案例 L 选择了过度概括的选项。其余八位参与者选择了有明确范围的选项，并指出案例 H 缺少语音证据，以及案例 L 未测量记忆、注意力或推理任务。他们建议的行动不是将单一结果转化为决策，而是记录情境、在更好的条件下重新采集数据，并比较多个样本或任务。

在群体层面，参与者对报告的依赖意愿因能力而异。对于总结单个样本（4 [2–5]）、标记缺失或质量较低的数据（5 [1–5]）以及比较重复样本（4 [2–5]），中位数评分高于推断页面中未提供的信息或在没有证据的情况下依赖总体分数（两者均为 1 [1–5]）。然而，参与者的评分差异仍然很大。在每个案例中，感知到的证据重要性中位数增加了 1 [0–2]。由于后续评分是在完整案例模块结束后进行的，这一变化反映了参与者对完整解读活动的响应，不能单独归因于报告页面。

5.3 跨报告包评析：汇总前的解释

自我报告的首次关注点有助于解释这些比较结果。仅评分报告（A套餐）的两位读者首先注意到较大的数字，解释优先报告（PackageC）的三位读者首先注意到前面的“该样本显示”或“无法显示”文本块，而五维报告（B套餐）的五位读者都通过数值概况进入。尽管这些并非视线轨迹，但这些报告展示了层级结构如何为同一证据提供不同的进入点。

看完所有报告套件后，在被问及哪一份报告最有可能使人只关注数字时，十人全部选择了仅评分报告（A套餐）。在被问及哪一份报告最容易引发担忧或最可能促使人过度反应时，十人也全部选择了该报告。P2将其中的69/100比作考试成绩。九人认为解释优先报告（Package C）最有助于支持正确理解。八人更喜欢将其用于参与者，因为范围、局限性和用途都先于指数呈现。P8和P10则更看重五维报告（B套餐）中可供检查的维度和证据，另有三人认为B套餐在质量和不确定性方面最清晰。

大多数参与者将报告解读为对当前样本的描述；与缺乏依据的推理相比，他们对总结观察到的证据给予了更高的信任评分。参与者更倾向于先看解释、再看分数，不过，也有一些人在查看内容更详细的报告后，对整体认知能力作出结论。

5.4 将两个边界转化为设计要求

首屏检查清单区分了参与者希望立即看到的信息，以及他们愿意按需查看的证据。9人选择了单样本声明、聚合值的名称与范围，以及数据质量；8人选择了通俗语言解读，以及缺失或不可用的证据。7人选择了警示，即看起来较低的结果无法补足未测量的认知能力；4人则要求在首屏展示全部五个维度及其来源。这种模式倾向于采用渐进式披露：边界与质量信息应始终紧邻任何汇总内容，同时让详细的来源信息易于查阅，但不主导初始阅读。

表7总结了设计启示。对于两个边界，界面必须揭示提示、复习和参与者的选择如何影响答案，以及汇总如何影响其解读。以人为中心的可解释性

表7。 基于*MindBridge*的设计启示 形成性研究。

设计重点	设计启示
后续提问措辞	后续提问应以明确陈述的细节为依据，避免无依据的假设，并保留与源轮次的链接。
跳过和完成控件	在响应字段旁始终显示带有“跳过”和“完成”标签的控件。
审阅和提示选择	记录审阅和提示选择，允许调整审阅时机，并避免重复提问。
评分呈现	在指数之前解释样本范围和局限性，并将其与指数并列显示。
证据详情	提供简明的初始解释，并允许展开查看维度和支持性证据。
缺失证据	区分不可用的证据与较低数值，并识别未被评估的能力或模态。
可读性与导航	支持调整文本大小、提供足够大的控件，并清晰标注证据访问入口。

因此，需要同时提供面向研究人员的记录和面向参与者的控件，将局限性紧邻相关数值呈现，并采用有利于界定范围的阅读方式的信息排序。

5.5 研究问题3：多模态分析管线的技术评估

为回答研究问题3，我们对*MindBridge*的分析流程进行了三项互补检查。首先，我们在固定录音上比较了不同模型配置，考察每种语料类别的总体准确率和错误情况。其次，我们检验了该流程能否根据文本和合成语音区分预先设定的沟通困难级别。第三，我们考察了同一案例在翻译和合成语言版本中的错误模式。这些评估刻画了分析流程的行为；它们并不能验证面向参与者的五维度摘要或研究信号指数是否能够作为认知状态的测量指标。附录B提供了固定录音清单、构建细节、操纵统计数据和跨语言协议。

固定录音模态审计。 我们分析了来自DementiaBank的109个英语Cookie Theft会话 [2]：其中48个为对照组录音，记为Corpus-C；另有61个录音标记为可能阿尔茨海默病、可能患有阿尔茨海默病或轻度认知障碍，合并为Corpus-NC，用于此次二分类审计。Corpus-NC作为正类；如果将每个录音都预测为该类别，准确率将达到56.0%。文本模型仅接收参与者发言轮次，而语音模型和语音加文本模型则可以接收完整录音，其中包括在106个会话中标注的研究人员语音。因此，输入范围与模态相互交织。表8报告了该固定清单上的配置级零样本准确率。

MindBridge 使用 GPT-5.4 进行转录文本分析和结构化证据融合，并使用 gpt-realtime-2 进行语音交互和定性语音分析。在固定录音评估中，这些模型在测试的纯文本配置（70.3%）和纯语音配置（66.1%）中，分别取得了最高准确率。这些结果反映的是各自单独的评估配置，而非二者组合后的部署表现。

然而，仅看准确率汇总值会掩盖重要的失效模式。在统一的开放模型协议下，Qwen2-Audio 和 Ultravox 对所有录音均预测为 Corpus-NC，Step-Audio 对 109 条录音中的 108 条作出了这一预测；Qwen2.5- Omni 的准确率为 62.4%，敏感度为 78.7%，特异度为 41.7%。这是一次失效模式审计，而非受控模态消融；它表明，汇总排名可能掩盖类别坍塌和访谈者语音暴露。

受控合成语音探测。 我们为 20 个中性老年人角色固定了一个六至九事件图，并将其以文本和合成语音的形式呈现为四个预设的沟通困难等级。全部八项文字记录和基于波形的操控测量指标均按预设方向发生变化（$p < .001$）。一份由盲测 Qwen2.5-Omni 完成的、提交至 ACM 的稿件

表8：固定的109条录音清单上单模态和多模态配置的准确率。† 表示 MindBridge 中使用的模型组件。GPT-5.4 也会进行证据融合。报告的准确率描述的是所评估的配置，而非其组合部署。

模型	模态	准确率（%）
单模态		
GPT-5.4†	Text	70.3
Qwen2.5-Omni-7B	Text	60.6
Qwen3-Omni-30B	Text	65.1
gpt-realtime-2†	语音	66.1
Audio-Flamingo-3	语音	44.0
Qwen2.5-Omni-7B	语音	55.0
Qwen3-Omni-30B	语音	56.9
多模态		
gpt-realtime-2	语音 + 文本	58.7
Audio-Flamingo-3	语音 + 文本	54.6
Qwen2-Audio-7B-Instruct	语音 + 文本	56.0
Qwen2.5-Omni-7B	语音 + 文本	62.4
Step-Audio-2-mini	语音 + 文本	56.9
Ultravox-v0.4.1-12B	语音 + 文本	56.0
Qwen3-Omni-30B	语音 + 文本	59.6

表 9。Qwen2.5-Omni-7B 的盲法序数敏感性结果。条件恢复指与编程设定的合成水平的一致性，而非对自然产生的人类样本的准确率。严格顺序指标将并列视为失败。QWK 表示二次加权 κ。

Mode	有效恢复		平均绝对误差 汇总 ρ		角色设定内 ρ		二次加权 Kappa	相邻严格	所有配对严格	完全严格
Text	80/80	**85.00%**	**.150**	**.940**	**.959**	**.939**	**83.33%**	**90.83%**	50.00%	
语音	80/80	26.25%	.975	.764	.867	.455	46.67%	62.50%	0.00%	
语音+文本	80/80	50.00%	.550	.837	.895	.691	58.33%	75.00%	0.00%	

探测测试在文本、语音及二者组合上完成了 240 次推理。表 9 报告了所得的序数敏感度测量结果。尽管对于 20 个角色中的 19 个，每种模态均保持了非递减顺序，但级别准确恢复率从文本的 85.0% 降至语音的 26.25% 和语音加文本的 50.0%。这些发现涉及对预设合成条件的区分，而非对自然产生的人类语音中的认知状态进行评估。

跨语言压力测试。 我们还将 109 个英语案例与中文和日语译文及合成语音配对，并保留了标注的言语不流畅现象。中文和日语的语音加文本准确率相同（62.39%），但敏感度/特异度分别从 52.46%/75.00% 变为 86.89%/31.25%。翻译、合成和语言同时发生了变化，因此该结果无法单独检验语言效应，也不能支持多语言可用性的结论。

这三项检查共同表明，不能仅凭总体准确率来描述*MindBridge*的分析流程。有些模型配置将所有录音都归为同一课程。在合成探测测试中，从语音中还原预设难度等级的准确度低于从文本中还原的准确度。翻译和合成的语言版本总体准确率也相同，但错误模式不同。

6 讨论

6.1 研究问题1：自适应访谈者是测量的一部分

研究问题1的发现表明，对自适应访谈者的评估不应只看其问题看起来有多相关，还应考察这些问题如何影响最终得到的证据。后续跟进以参与者的回答为依据，并不意味着可以省去对问题所引入假设的检查。

后续提问应保留与参与者所提供细节的关联，而跳过和完成控件应清晰可见，而不应依赖参与者知道如何在对话中拒绝。重读和提示重定向同样应作为任务条件记录，因为它们会改变刺激暴露以及所叙述的体验。保留这些背景信息有助于研究人员结合回答产生的方式来解读回答，而不是将每一项特征都归因于参与者。

6.2 研究问题2：结果界面是推断的一部分

研究问题2支持将先解释后汇总作为一种设计方向，而非普遍更优的界面。报告应在指数之前说明样本范围、数据质量和缺失证据，并按需提供维度和来源证据。这样可以将初步解读所需的信息，与检查其依据所需的详细信息区分开来。

清晰的呈现并不保证解读范围明确。每项评分的方向都需要在数值旁加以说明：诸如“话语流畅度”这样的标签，即使数值越高代表与困难相关的模式越明显，也可能让人误以为能力更强。因此，评估应考察读者自己的解释和预期用途，而不应仅凭他们对报告的偏好、信心或对提示问题的回答来推断其是否理解。

6.3 研究问题3：分析流程是证据链的一部分

研究问题3表明，解读输出时应同时考虑生成输出时所处的条件。由于模型配置和输入范围同时发生变化，固定录音的比较无法单独确定模态效应。翻译和语音合成也使语言效应无法单独确定。

对研究人员而言，工作台应将输出与模型配置、输入来源、模态、语言及缺失证据一并呈现，并在检查评估结果时提供查看各数据集类别错误的途径。对参与者而言，报告应指出实际使用了哪些证据，同时避免暗示所有模态的贡献都相同。这样的区分有助于详细调查分析流程中的故障，而无需参与者解读基准诊断。

6.4 贯穿证据流程的以人为中心的可解释性

对于研究问题1–研究问题3，可解释性涉及如何获取、分析和呈现认知监测证据（图 1）。特征级解释不能替代对提示和任务条件的记录，而完整的对话日志也无法确保报告传达其主张的局限性。

对于 *MindBridge*，其设计目标是将每项报告主张与支持性观察、用于推导这些观察的输入，以及收集这些输入时所处的任务条件关联起来。这使参与者控制与研究者检查形成互补：前者让参与者能够在数据收集过程中作出选择；后者则揭示这些选择及后续分析如何塑造结果。同一原则也可用于自适应访谈、健康信息采集代理和教育稿件提交至 ACM

或同时引导并评估响应的反思系统。可追溯性有助于审查，但其本身并不能证明某项判断是有效的。

6.5 主张范围与后续步骤

这些发现为认知监测的设计提供了依据，但并不能确立临床有效性。临床意见来自一名精神科住院医师，而这项规模较小、年龄混合的用户研究考察的是文字交互，而非语音使用或记忆表现。整套软件包的比较并不能确立界面因果效应或年龄差异。技术检查受到语料库构成、不同输入范围以及翻译或合成语音的限制。附录D详细说明了这些限制。

后续评估应纳入更广泛的目标用户样本，并使用语音，考察麦克风状态清晰度、延迟、轮次结束控制和转录修正。应将行为记录与开放式解释配对，以评估参与者是否能够控制数据收集，并理解其结果所支持的内容。还需要对语音证据进行独立表征。

7 结论

MindBridge 是一款交互式、多语言、多模态认知监测工具，通过多模态分析连接两个面向人的流程：访谈者协助形成证据，分析过程可以保留或压缩证据，而结果界面则帮助解读证据。形成性研究将这些流程转化为具体要求：后续提问应以参与者提供的锚点为依据；跳过和完成控件应始终显示；后续提问、刺激材料复习和提示选择应保留在各自对应的记录中；在任何汇总值之前及其旁边，都应呈现样本范围、质量和缺失性。固定录音和合成审计表明，为何必须确保模态、语言和类别条件行为仍可供检查。因此，对自然口语交互开展更大规模的研究，是检验这一交互式、可解释认知监测方法的下一项必要工作。

参考文献

[1] Saleema Amershi、Dan Weld、Mihaela Vorvoreanu、Adam Fourney、Besmira Nushi、Penny Collisson、Jina Suh、Shamsi Iqbal、Paul N. Bennett、Kori Inkpen、Jaime Teevan、Ruth Kikin-Gil 和 Eric Horvitz。2019。《人机交互指南》。载于计算系统中的人为因素*CHI*会议*2019*论文集。计算机协会，纽约，美国纽约州，文章编号3，共13页。doi:10.1145/3290605.3300233[2] James T. Becker、François Boller、Oscar L. Lopez、Judith Saxton 和 Karen L. McGonigle。1994。《阿尔茨海默病的自然病程：研究队列描述与诊断准确性》。《神经病学档案》 51卷6期（1994），585–594。doi:10.1001/archneur.1994.00540180063015[3] Veronica Boschi、Eleonora Catricalà、Monica Consonni、Cristiano Chesi、Andrea Moro 和 Stefano F. Cappa。2017。《神经退行性语言障碍中的连贯语音：综述》。《心理学前沿》 8卷（2017），269。doi:10.3389/fpsyg.2017.00269[4] Zana Buçinca、Maja Barbara Malaya 和 Krzysztof Z. Gajos。2021。《信任还是思考：认知强制功能可减少人工智能辅助决策中的人工智能过度依赖》。《ACM人机交互汇刊》 5卷，CSCW1，文章编号188（2021），共21页。doi:10.1145/3449287[5] Canary Speech。2024。Canary Cognitive。https://canaryspeech.com/solutions/segment/cognitive/[6] Sofia de la Fuente Garcia、Craig W. Ritchie 和 Saturnino Luz。2020。《用于阿尔茨海默病监测的人工智能、语音与语言处理方法：系统综述》。《阿尔茨海默病杂志》 78卷4期（2020），1547–1574。doi:10.3233/JAD-200888[7] Zijian Ding、Jiawen Kang、Tinky Oi Ting Ho、Ka Ho Wong、Helene H. Fung、Helen Meng 和 Xiaojuan Ma。2022。《TalkTive：利用反馈语让老年人参与神经认知障碍筛查的对话代理》。载于计算系统中的人为因素*CHI*会议*2022*论文集。计算机协会，纽约，美国纽约州，文章编号304，共19页。doi:10.1145/3491102.3502005[8] Zhihao Du、Changfeng Gao、Yuxuan Wang、Fan Yu、Tianyu Zhao、Hao Wang、Xiang Lv、Hui Wang、Chongjia Ni、Xian Shi、Keyu An、Guanrou Yang、Yabin Li、Yanni Chen、Zhifu Gao、Qian Chen、Yue Gu、Mengzhe Chen、Yafeng Chen、Shiliang Zhang、Wen Wang 和 Jieping Ye。2025。《CosyVoice 3：通过扩展规模与后训练实现真实场景语音生成》。arXiv:2505.17589。doi:10.48550/arXiv.2505.17589[9] Marshal F. Folstein、Susan E. Folstein 和 Paul R. McHugh。1975。《“简易精神状态检查”：临床医生评定患者认知状态的实用方法》。《精神病学研究杂志》 12卷3期（1975），189–198。doi:10.1016/0022-3956(75)90026-6[10] Kathleen C. Fraser、Jed A. Meltzer 和 Frank Rudzicz。2016。《语言学特征可从叙事语音中识别阿尔茨海默病》。《阿尔茨海默病杂志》 49卷2期（2016），407–422。doi:10.3233/JAD-150520

[11] FRONTEO公司。2015。KIBIT。https://www.fronteo.com/kibit[12] Jonathan Heitz、Gerold Schneider 和 Nicolas Langer。2025。GPT-4提取的语言学特征可提升基于自发性言语的阿尔茨海默病检测效果。载于第*31*届国际计算语言学会议论文集。计算语言学协会，阿布扎比，阿拉伯联合酋长国，1850–1864。https://aclanthology.org/2025.coling-main.126/[13] Toshiro Horigome、Kimihiro Hino、Hiroyoshi Toyoshiba、Norihisa Shindo、Kei Funaki、Yoko Eguchi、Momoko Kitazawa、Takanori Fujita、Masaru Mimura 和 Taishiro Kishimoto。2022。使用自由对话的向量表示识别神经认知障碍。《科学报告》12，文章编号12461（2022），8页。doi:10.1038/s41598-022-16204-4[14] Jiaxiong Hu、Junze Li、Yuhang Zeng、Dongjie Yang、Danxuan Liang、Helen Meng 和 Xiaojuan Ma。2024。为神经认知障碍筛查对话任务中的对话代理设计支架策略。载于*2024*年计算机系统人为因素*CHI*会议论文集。计算机协会，纽约，美国，文章编号70，共21页。doi:10.1145/3613904.3642960[15] Jessica Hullman、Xiaoli Qiao、Michael Correll、Alex Kale 和 Matthew Kay。2019。探寻误差：不确定性可视化评估调查。《*IEEE*可视化与计算机图形学汇刊》25，1（2019），903–913。doi:10.1109/TVCG.2018.2864889[16] Matthew Kay、Tara Kola、Jessica R. Hullman 和 Sean A. Munson。2016。我的公交车什么时候（大概）到？日常移动预测系统中以用户为中心的不确定性可视化。载于*2016*年计算系统中的人为因素*CHI*会议论文集。计算机协会，纽约，美国，5092–5103。doi:10.1145/2858036.2858558[17] Q. Vera Liao、Daniel Gruen 和 Sarah Miller。2020。质疑人工智能：为可解释人工智能用户体验提供设计实践参考。载于*2020*年计算系统中的人为因素*CHI*会议论文集。计算机协会，纽约，美国，1–15。doi:10.1145/3313831.3376590[18] Isaac M. Lipkus。2007。传达健康风险的数字、文字和视觉格式：最佳实践建议与未来展望。《医学决策》27，5（2007），696–713。doi:10.1177/0272989X07307271[19] Kimberly D. Mueller、Bruce Hermann、Jonilda Mecollari 和 Lyn S. Turkstra。2018。轻度认知障碍和阿尔茨海默病中的连贯言语与语言：图片描述任务综述。《临床与实验神经心理学杂志》40，9（2018），917–939。doi:10.1080/13803395.2018.1446513[20] Ziad S. Nasreddine、Natalie A. Phillips、Valérie Bédirian、Simon Charbonneau、Victor Whitehead、Isabelle Collin、Jeffrey L. Cummings 和 Howard Chertkow。2005。蒙特利尔认知评估（MoCA）：轻度认知障碍的简短筛查工具。《美国老年医学会杂志》53，4（2005），695–699。doi:10.1111/j.1532-5415.2005.53221.x[21] Paula Andrea Pérez-Toro、Sebastian P. Bayerl、Tomás Arias-Vergara、Juan Camilo Vásquez-Correa、Philipp Klumpp、Maria Schuster、Elmar Nöth、Juan Rafael Orozco-Arroyave 和 Korbinian Riedhammer。2021。ADReSSo挑战赛背景下访谈者对阿尔茨海默病自动评估的影响。载于*Interspeech 2021*论文集。国际语音通信协会，布尔诺，捷克，3785–3789。doi:10.21437/Interspeech.2021-1589[22] Ulla Petti、Simon Baker 和 Anna Korhonen。2020。基于语音与语言的阿尔茨海默病自动检测系统性文献综述。《美国医学信息学会杂志》27，11（2020），1784–1797。doi:10.1093/jamia/ocaa174[23] Jessica Robin、Mengdan Xu、Liam D. Kaufman 和 William Simpson。2021。利用数字言语评估检测认知障碍的早期迹象。《数字健康前沿》3（2021），749758。doi:10.3389/fdgth.2021.749758[24] Fatemeh Taherinezhad、Mohamad Javad Momeni Nezhad、Sepehr Karimi、Sina Rashidi、Ali Zolnour、Maryam Dadkhah、Yasaman Haghbin、Hossein Azadmaleki 和 Maryam Zolnoori。2026。基于语音的认知筛查中的大语言模型适配策略：系统评估。*JMIRAI* 5（2026），e82608。doi:10.2196/82608[25] Vanessa Taler 和 Natalie A. Phillips。2008。阿尔茨海默病与轻度认知障碍中的语言表现：比较综述。《临床与实验神经心理学杂志》30，5（2008），501–556。doi:10.1080/13803390701550128[26] Johannes Tröger、Ebru Baykara、Jian Zhao、Daphne ter Huurne、Nina Possemis、Elisa Mallick、Simona Schäfer、Louisa Schwed、Mario Mina、Nicklas Linz、Inez Ramakers 和 Craig Ritchie。2022。远程自动化ki:e轻度认知障碍认知言语生物标志物的验证：依据DiME V3框架进行验证与确认。《数字生物标志物》6，3（2022），107–116。doi:10.1159/000526471[27] Ziang Xiao、Michelle X. Zhou、Q. Vera Liao、Gloria Mark、Changyan Chi、Wenxi Chen 和 Huahai Yang。2020。告诉我关于你自己的事：使用人工智能驱动的聊天机器人开展包含开放式问题的对话式调查。《*ACM*计算机与人类交互汇刊》27，3，文章编号15（2020），37页。doi:10.1145/3381804

附加系统界面截图

以下屏幕展示了正文所述交互与溯源机制在整页视图中的呈现方式。面向参与者的界面集包含形成性研究中使用的有界追问界面（通过键入作答），以及两个互补的任务界面。这些屏幕共同展示了参与者在哪里修改或提交回答、复习阅读刺激材料、调整自传式提示，或跳过任务。因此，它们将塑造引 elicited 记录的控件清晰呈现出来，而不只是将任务作为流程说明来介绍。

面向研究者的界面集通过一个合成演示会话展示证据链的另一面。报告指出来源模态、质量、缺失证据和潜在混杂因素。对话视图按顺序展示该会话中的轮次，并标明说话者角色和输入模式。这些屏幕说明了面向参与者的控制与面向研究者的可追溯性之间预期的联系；它们属于界面文档，而非额外的研究结果。所有显示的编号、对话和结果均为虚构或合成内容，并未展示参与者数据。

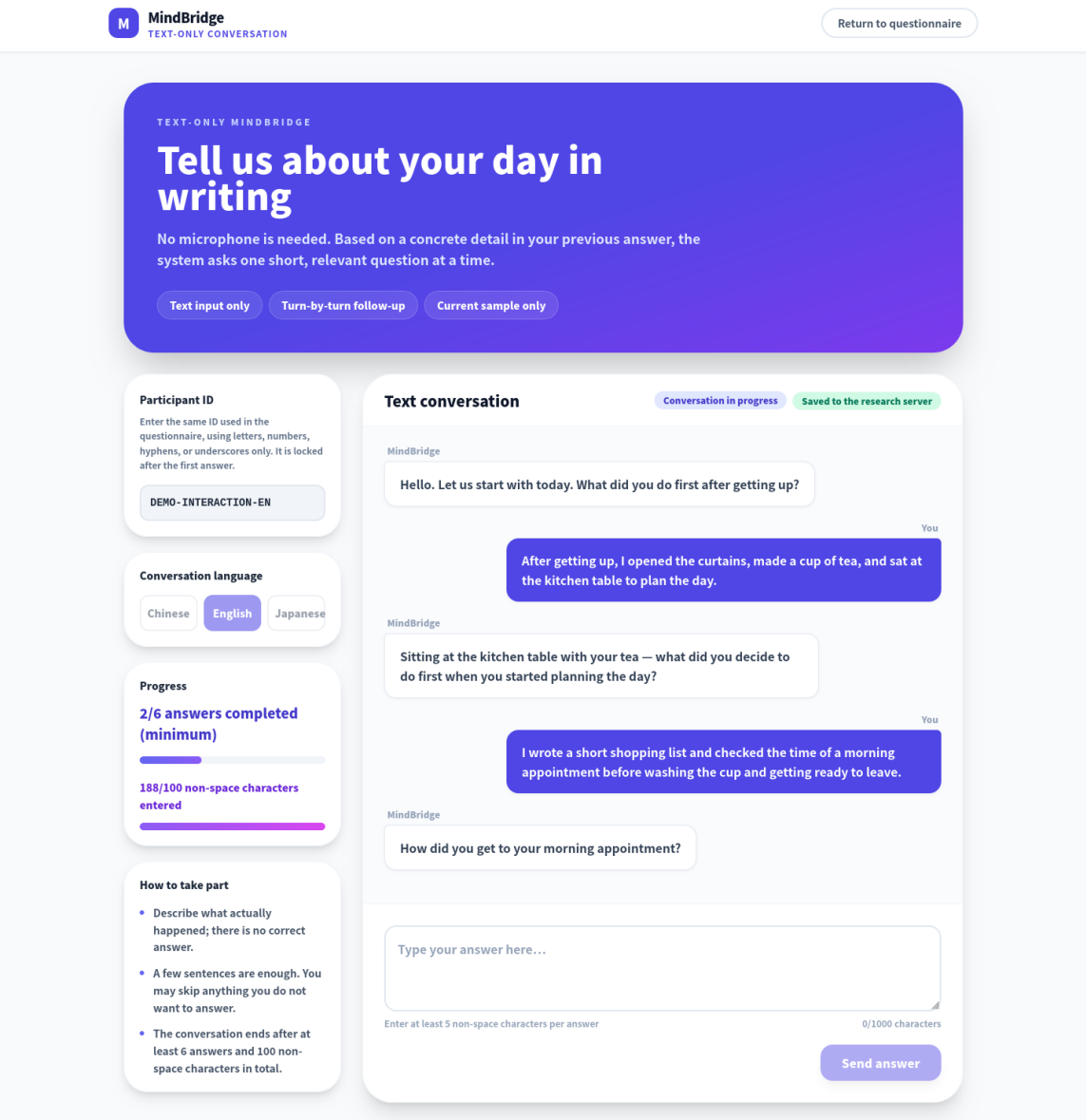


图 5。研究专用的类型化有限追问界面。左栏展示参与者背景、语言、进度和交互指导；对话面板保留有序轮次，并让参与者控制修订和提交。

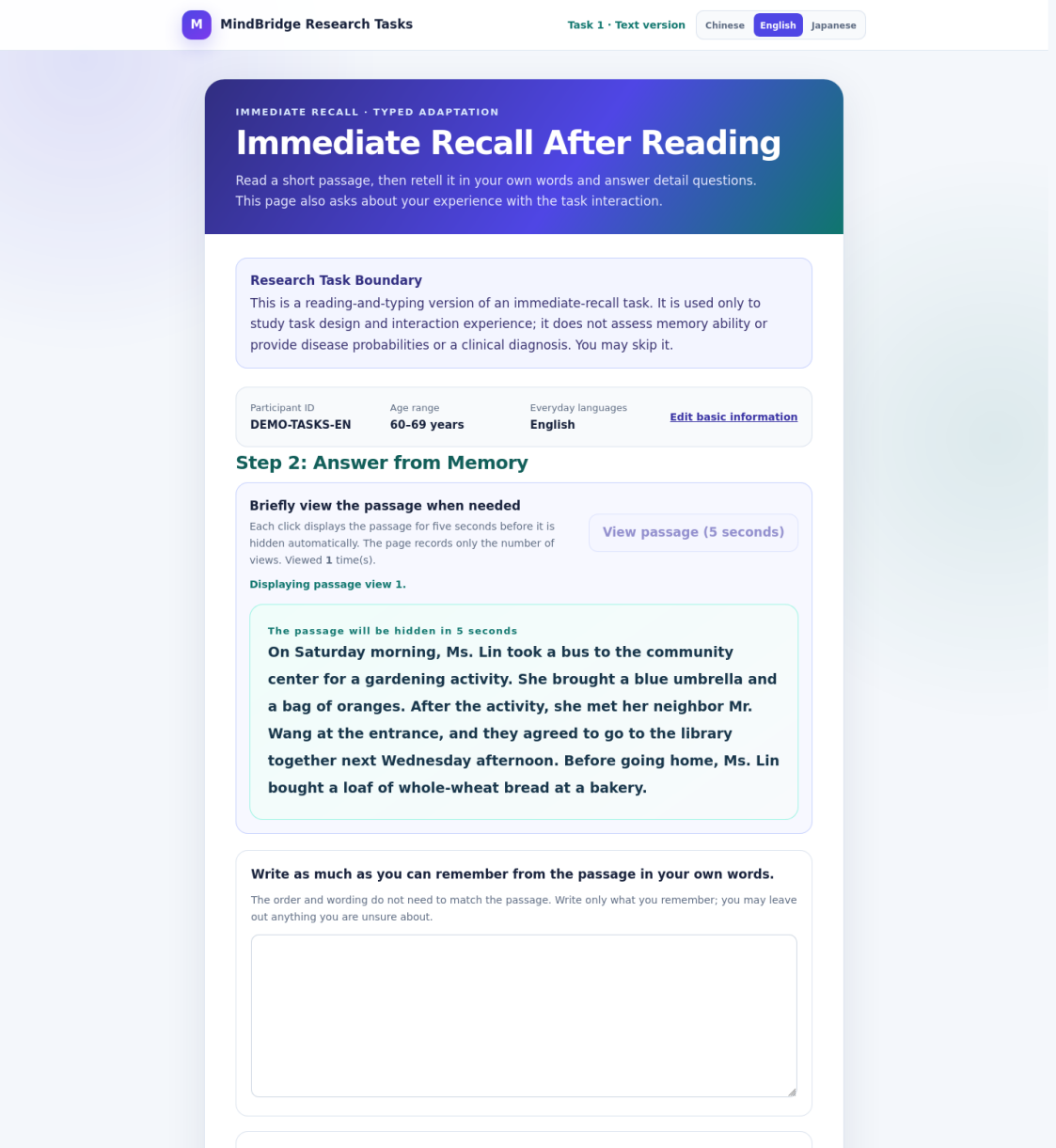


图 6. 阅读后立即回忆任务。界面将简短的源文段、可选的五秒复习、参与者背景和自由回忆答题区同时显示。

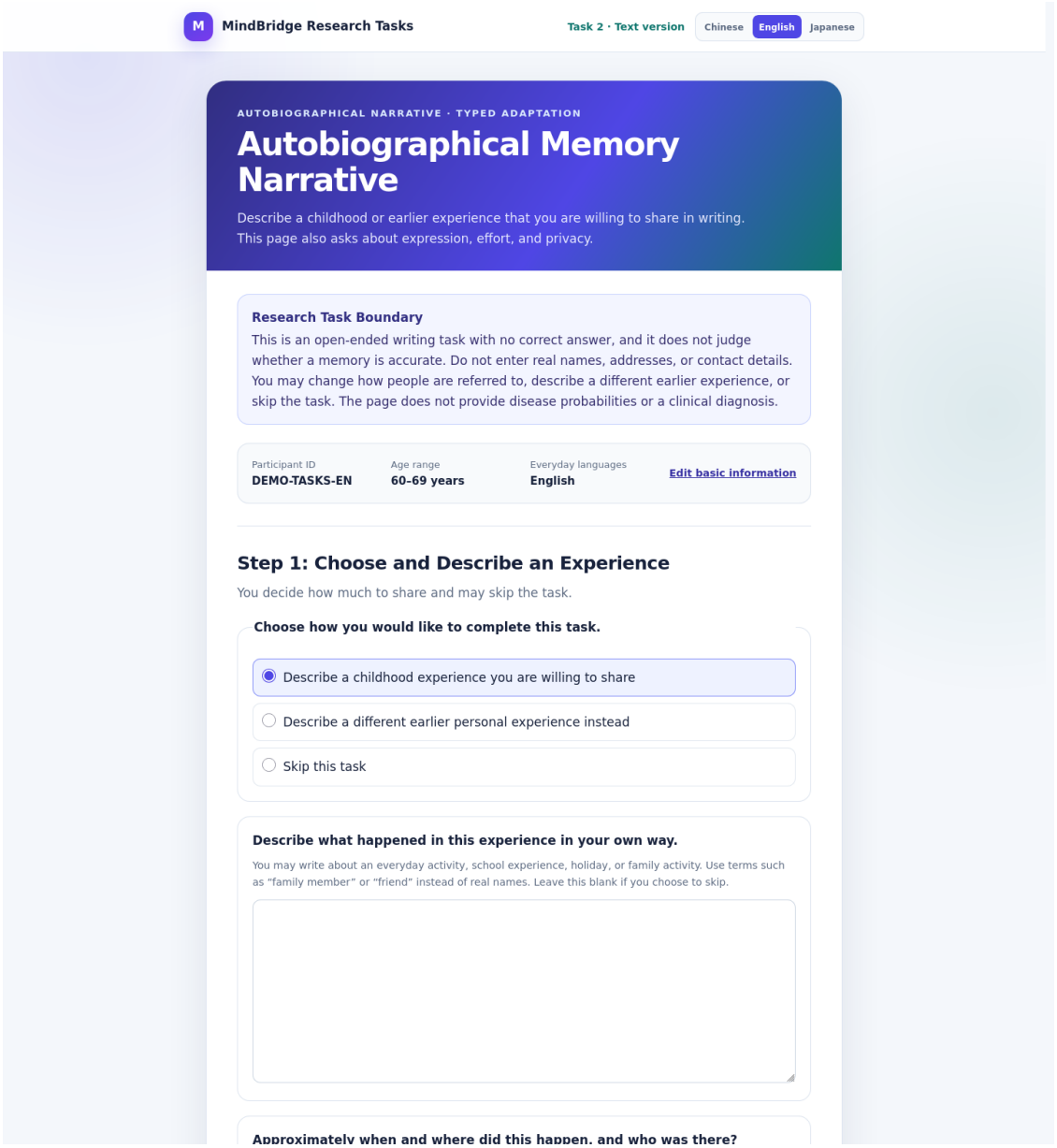


图 7. 自传式记忆叙述任务。参与者可选择一段童年经历、另一段更早的个人经历，或选择明确跳过，然后决定以书面形式披露哪些内容。

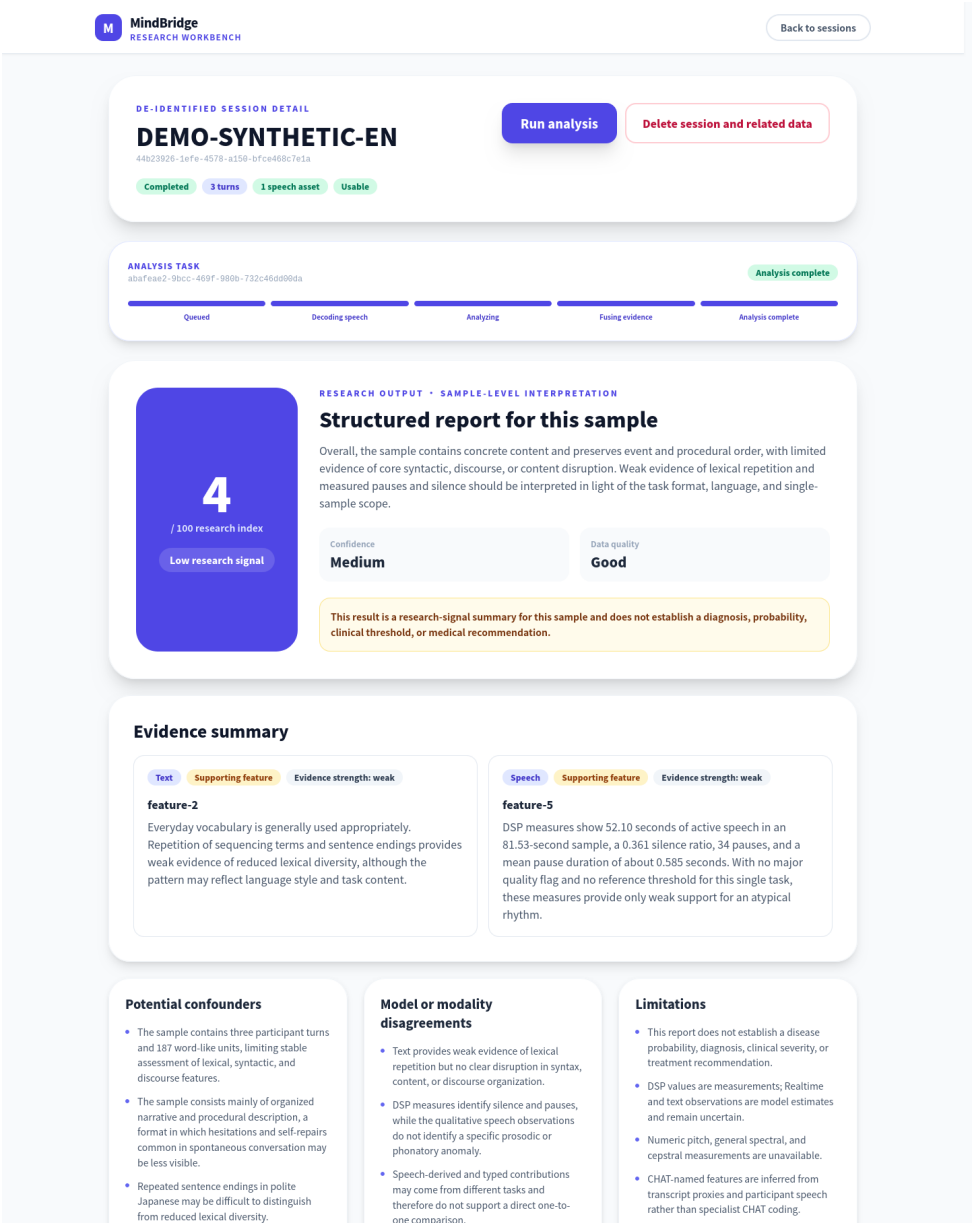


图 8. 面向研究人员的合成演示会话报告视图。英语翻译呈现了对日语样本的解读、数据质量、按来源模态标注的观察结果、潜在混杂因素、模态分歧和局限性。

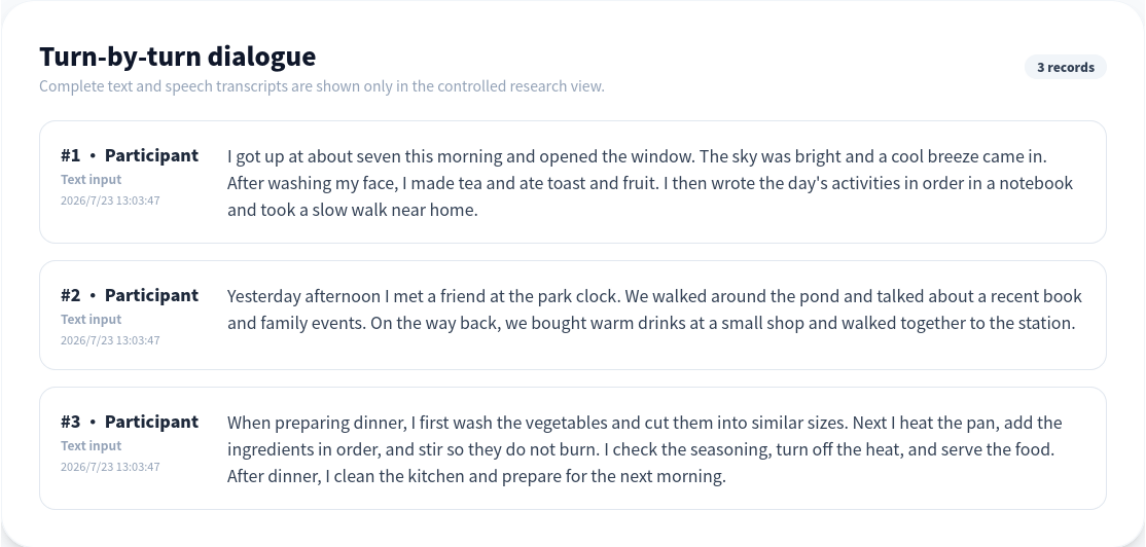


图 9。排序后的对话视图，展示同一合成演示会话。三轮参与者发言以英语翻译呈现及其顺序、角色、输入模式和时间戳。

B 技术边界检查详情

B.1 固定录音清单与输入范围

固定录音审计使用了DementiaBank中原始、未经降噪处理的英语 Pitt 语料库里的Cookie Theft图片描述会话 [2], ，并将每段录音与其CHAT转录文本配对。[1] 每次运行都使用同一份预先存在的输入列表作为固定的评估清单，而非训练测试划分；本文不包含特定任务的训练实验。

表 10 总结了109场会话。CHAT组别字段显示，其中48场属于对照组；其余61场包括50场可能阿尔茨海默病、4场可能患有阿尔茨海默病和7场轻度认知障碍。我们仅在二元评估中合并后面这些类别，并以中性的名称Corpus-C和Corpus-NC报告目标类别。要求输出的字面标签为Control和Impaired；这些标签沿用了本次评估所使用的源语料类别。

这些会话对应97个参与者标识符；其中11个重复出现，但没有一个跨越合并后的标签类别。评估指标以录音为分析单位。109个可读取的MP3文件总时长为123.35分钟，每个配置均对其进行了预测。纯文本模型接收的清理后的转录文本仅包含参与者（PAR）话轮，而语音模型接收完整的会话录音，语音加文本模型则同时接收两者。106份转录文本中至少标注了一段调查者（INV）话轮。Corpus-C中调查者话轮的中位数为3，Corpus-NC中为5；相应的标注的调查者语音时长中位数分别为2.1秒和5.9秒。因此，语音分支可能会暴露文本分支无法获取的访谈者行为，从而形成一种捷径，而非仅基于参与者的证据。

B.2 固定录音模型比较

正文中的表8报告了所有受评单模态和多模态配置的准确率。

零样本 指不使用特定任务微调示例的固定指令，但这并不保证预训练过程中未使用 Demen- tiaBank。由于架构和输入范围均发生变化，这属于失效模式审计，而非受控模态消融，也不能作为多模态更优的证据。在通用开放模型多模态协议下，Qwen2-Audio 和 Ultravox 对全部 109 段录音均预测为 Corpus-NC（敏感度为 100%，特异度为 0%）；Step-Audio 对 109 段录音中的 108 段预测为 Corpus-NC（敏感度为 100%，特异度为 2.1%）。Qwen2.5-Omni 的准确率为 62.4%，敏感度为 78.7%，特异度为 41.7%。因此，仅凭汇总准确率无法确定界面内容或生产模型选择。

B.3 受控合成语音构建

我们根据人口统计属性、职业、地区和日常生活锚点，构建了 20 个中性老年人角色画像。角色设定中不包含有关认知状态或预设难度的标签或提示。这组角色包括 10 名女性和 10 名男性，年龄在 66 至 87 岁之间，受教育年限为 10 至 18 年；职业和地区均各不相同。10 个角色使用晨间日常提示，另有 10 个使用职业特定提示。

对于每种角色设定，GPT-4o 都生成一个包含六至九个底层事件的固定图结构，并在四种预设的沟通困难等级下呈现相同结构。S0 表示流畅、具体且信息密集；S1 加入偶尔的犹豫和模糊表达；S2 则增加频繁的找词困难、重复、修复、填充性停顿和话语中断；而

[1]英语 Pitt 语料库：https://talkbank.org/dementia/access/English/Pitt-orig.html；语料库 DOI：https://doi.org/10.21415/T5CS3J。我们感谢 N　IA 资助 AG03705　和 AG05133，这些资助支持了 Pitt 语料库的数据收集。

表 10。基于 fromCHAT 元数据的固定英语评估清单构成。除参与者 ID 外，计数均为录音级。

衡量	All	Corpus-C	Corpus-NC
会话录音	109	48	61
参与者 ID	97	40	57
年龄，中位数（范围）	69（48–90）	66（48–80）	74（53–90）
性别字段（F/M）	75/34	32/16	43/18
语音时长（分钟）	123.35	48.41	74.94

表 11。 Mea每个设定级别下 20 个角色的 n（SD）语言和声学测量。填充词、省略号、模糊表达 s, 以及修复或 l ost-thread 标记为每 100 个词语的比率。TTR 是类符形符比；Dur. 是以秒为单位的时长。

级别	词语	TTR	填充词	省略号	模糊表达	修复	Dur.	静默	停顿
S0	116.70(11.94)	.681(.049)	0.000(0.000)	0.000(0.000)	.294(.486)	0.000(0.000)	40.23(4.81)	.313(.026)	19.95(4.25)
S1	104.40(11.20)	.652(.055)	6.747(2.368)	2.075(1.553)	.571(.707)	0.000(0.000)	46.37(5.58)	.394(.023)	29.95(5.41)
S2	108.55(11.10)	.576(.043)	12.690(2.765)	9.260(3.694)	1.553(1.551)	.277(.437)	71.39(13.15)	.524(.044)	48.50(9.98)
S3	114.20(10.46)	.510(.048)	20.083(4.446)	21.393(6.179)	1.823(1.502)	.764(.892)	114.26(22.95)	.646(.061)	70.10(12.22)

S3 增加了普遍存在的犹豫、未解决的遣词困难、中途放弃的言语、思路中断以及持续性重复。主题内容和预期事件保留在角色设定中，而语言表达方式发生了变化。

我们使用 gpt-4o-mini-tts，采用 coral 语音预设和针对不同级别的语音呈现指令，合成每种呈现形式。对于 S1–S3，实现方案在句子边界和省略号处拆分文本，对各片段进行合成，并在片段之间插入本地静默后进行拼接；静默时长随级别升高而增加。该材料包含 80 份转录文本，以及 80 个单声道、24 kHz 的 WAV 文件，总时长为 90.75 分钟。

B.4 合成操控与序数敏感度检查

根据逐字转录稿统计表层语言标记，并按每 100 个词进行归一化。静音比率和停顿次数根据最终 WAV 文件，使用帧级均方根阈值重新计算；它们是操作性信号指标，而非经人工验证的声学标注。表 11 展示了词汇多样性、言语不流畅和语音时序如何随四个设定级别而变化。

重复测量的弗里德曼检验将角色设定作为重复测量单位。所有八项指标都按设定方向变化（$p < .001$）；肯德尔W的范围为：模糊用语对应的 .357 至静音比率对应的 1.000。这支持内部操控控制，但不能证明其与人类损伤相对应。

Qwen2.5-Omni-7B 用作敏感度探测测试，而非真实标注者。它完成了 240 次确定性零样本推理：20 个角色 × 四个级别 × 三种输入模式。每项刺激材料均在独立上下文中呈现，并采用固定的随机顺序；角色元数据和设定级别均予以隐藏。提示要求仅对可观察到的沟通困难进行 0–3 评分，并禁止得出超出刺激材料范围的结论。

正文中的表 9 报告了三种输入模式下完整的序数敏感性结果。

在 20 个角色中，有 19 个在所有模态下都保持了四级非递减顺序，但只有文本能让相当一部分角色的四个级别都严格区分开来（50%；包含语音时为 0%）。语音评分平均比文本低 1.05 分（角色自助法 95% 置信区间 [-1.175, -.938]），且包含语音时，探测测试从未给出级别 3。造成这种情况的可能是模型的语音路径、文本转语音的实现，或两者的交互作用。整个过程中使用了相同的语音预设；S0 的录音异常清晰，而且性别与任务相互混杂。

B.5 跨语言压力测试方案

我们将 109 个英语案例与中文和日语译文配对，译文保留了标注的停顿、重复、修正、片段和非语言事件。CosyVoice3 使用每段源录音的八秒参考窗口来生成译文语音 [8]。这样既保持了案例身份和类别平衡，又同时改变了语言、翻译和合成。

Qwen2.5-Omni-7B 对中文和日语产生了相同的语音加文本准确率（62.39%），但敏感度/特异度分别为 52.46%/75.00% 和 86.89%/31.25%。尽管中文的敏感度/特异度为 50.82%/66.67%，日语的相应数值为 96.72%/8.33%，纯语音准确率同样保持一致（57.80%）。这种变化可能源于翻译后的言语不流畅现象、合成韵律、模型语言先验或它们之间的交互；该检查无法单独分离出语言效应。

C 面向参与者的评分构建

GPT-5.4 证据融合组件通过固定的结构化提示，按照预定义的评分方向，为每个保留特征分配一个从 0 到 100 的整数评分 s_i。该提示将 0 定义为在来源可用时未发现明显匹配模式，将 1–24 定义为轻微或孤立，将 25–49 定义为轻度，将 50–74 定义为明确，将 75–100 定义为突出或普遍存在。允许使用这些锚点之间的数值，且每个评分都必须附带一项特定来源的观察和置信度判断。这些锚点用于指导模型输出，并非经过验证的临床阈值。对于可用特征集合 A，可信服务器计算某个维度或总体分数，公式如下：

$$S = \mathrm{round}\left(\frac{\sum_{i \in A} w_i s_i}{\sum_{i \in A} w_i}\right),$$

其中，$w_i = 1$ 表示核心特征，$w_i = 0.5$ 表示辅助特征。维度计算使用该维度内可用的特征。总体指数直接对所有可用的保留特征应用相同规则，而不是对五个维度的分数取平均。缺失特征会从两个求和项中排除；如果没有可用特征，则不生成该指数。

这些权重体现的是医生的序数优先级，而不是预测价值的测量差异。固定的 2:1 比率使每个核心特征的权重为每个辅助特征的两倍。我们未评估其他加权比率。

该界面将 0 至 33 的总体值标记为低模式显著程度，34 至 66 标记为中等，67 至 100 标记为高模式显著程度。这些是描述性的显示区间，并非经过验证的临床阈值。

D 详细局限性

DementiaBank 审计描述的是一份固定的英语 Cookie Theft 清单，而非随机样本或与总体匹配的样本。其中的 109 个会话对应 97 个参与者标识符，因此重复访问并不代表彼此独立的人员；所有评估指标均为录音级，无法确立参与者级或总体级表现。Corpus-C 和 Corpus-NC 在年龄、访谈者贡献以及部分媒体格式方面也存在描述性差异。文本条件仅包含参与者发言轮次，而完整会话语音可能包含研究人员语音；因此，模态配置之间的差异不能作因果解释。零样本描述的是我们的仅提示词流程，而不是模型预训练；我们也无法排除此前接触过 DementiaBank 的可能性。

*MindBridge*的临床依据来自一名精神科住院医师在多个阶段的参与：初次咨询、针对11类清单作出的33项结构化判断、复习带标签的样本语音，以及书面

说明特征优先级和披露的依据，并复习原型用户界面。这不构成多临床医生共识、参与式协同设计或临床验证。我们汇编了清单，将保留的类别整理为五个维度，并将类别判断转换为数值权重。由此得到的0–100研究信号既未标准化，也未参照常模，因此无法确定总体认知能力。

用户研究采用了一个规模较小、由十人组成的混合年龄便利样本，其中年龄段为 60 岁及以上的参与者仅有三人；参与者首次接触仅评分报告（A套餐， $n = 2$）、五维报告（B套餐，$n = 5$）和解释优先报告（C套餐，$n = 3$）的情况也不均衡。P9 针对高指数案例（案例 H）查看了 B套餐，针对低指数案例（案例 L）查看了 C套餐；其余九名参与者则在两个案例中始终使用同一套餐。该样本无法估计年龄效应、人群患病率或报告套件的优越性。已披露的单样本边界、先固定查看案例 H 再查看案例 L 的顺序、P9 接触内容的异质性，以及之后对所有套餐的比较，都会引入指导与迁移效应。首次关注和适应情况由参与者自我报告；研究专用条目并非经验证的量表；开放式回答编码也未经独立的双重编码。两项互补任务仅使用一篇文章和一个开放式叙事提示，其评分评估的是交互，而非记忆准确性。导出数据缺少招募、引导、设备、实际时长和构建版本元数据。打字无法评估麦克风控件、自动语音识别、言语时序、声学特征或语音交互可用性。

合成扩展使用的是生成语言和语音，而非经过独立表征的人类语音。包含 20 种角色设定的批次采用了单一文本转语音语音预设，存在任务—性别混淆因素，并实施了固定的单轮协议。即使使用不同模型，仍可能存在生成器—评估器循环性。跨语言检查结合了翻译和语音合成，因此无法分离语言效应，也无法估计自然生成的中文或日语语音上的表现。这些限制界定了后续评估自然生成的多语言语音及其在真实场景中的应用所需的证据。