arXiv:2602.01387v1 [cs.HC] 1 Feb 2026

谨慎披露：设计访谈聊天机器人的隐私控制

Ziwen Li
li.ziw@northeastern.edu 东北大学 波士顿，马萨诸塞州，美国

Ziang Xiao
ziang.xiao@jhu.edu 约翰斯·霍普金斯大学 巴尔的摩，马里兰州，美国

Tianshi Li
tia.li@northeastern.edu 东北大学 波士顿，马萨诸塞州，美国

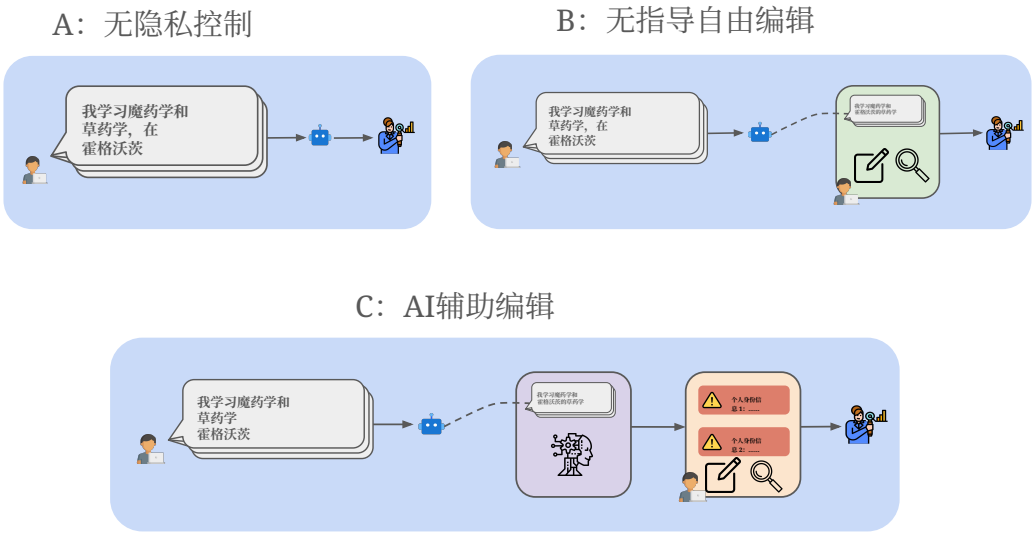

图 1：我们考察了三种用于研究访谈聊天机器人隐私控制的实验条件，探究其对隐私保护和数据质量的影响。(A) 无隐私控制（基线）：聊天机器人未经干预，直接将参与者的原始回答转发给研究人员，既保留了数据的真实性，也未提供任何保护措施。(B) 无指导自由编辑：访谈结束后，提示参与者自行查看并编辑对话日志，赋予其自主权，但不提供结构化指导。(C) AI辅助编辑：大语言模型（LLM）自动检测并标记潜在的个人身份信息（PII），通过指导参与者进行访谈后编辑，在降低隐私泄露风险的同时保留用户控制权。

摘要

在 HCI 中，收集敏感话题相关的数据仍然具有挑战性，因为参与者常常会由于隐私顾虑和社会赞许偏差而隐瞒信息。尽管聊天机器人被感知的匿名化特征可能会减少这些障碍，但研究却反常地表明，人们倾向于向聊天机器人过度分享个人信息或敏感信息。在本研究中，我们探索了聊天机器人访谈中的隐私控制，以应对这一问题。该隐私控制允许参与者在访谈结束时修改其访谈记录，并提供两种设计变体：自由编辑和 AI辅助编辑。在一项被试间研究中（$N = 188$），我们比较了在一个关于敏感话题的基于聊天机器人的访谈中，无编辑、自由编辑和 AI辅助编辑条件。结果证实了基于聊天机器人的访谈中普遍存在过度分享问题，并表明 AI辅助编辑是一种有效的隐私控制机制，能够减少个人身份信息披露同时保持数据质量和用户参与度，从而为在此类访谈中平衡伦理实践与数据质量提供了一种有前景的方法。

CCS 概念

• 安全与隐私 → 安全与隐私的人文和社会层面；• 以人为中心的计算 → 人机交互；• 计算方法 → 话语、对话与语用学。

关键词

大语言模型（LLM）、通用人工智能（AGI）、对话代理、聊天机器人、个人数据/追踪、隐私、会话分析、访谈、隐私增强技术

1 引言

在人机交互（HCI）和社会科学研究中，参与者倾向于隐瞒或歪曲信息，这往往会损害敏感话题相关数据的质量 [50, 51]。这一现象通常源于对隐私的担忧 [20]、社会赞许偏差 [8, 30]，或对潜在后果的顾虑 [64]，会直接损害自我报告数据的效度和信度。当人们掩饰自己在健康状况、财务行为或个人信念等敏感话题上的真实看法、行为或经历时，研究人员便无法全面或准确地了解所研究的现象 [7, 16]。因此，此类研究得出的见解会存在偏差，进而形成有缺陷的理论模型、无效的设计建议和不当的干预措施，既无法应对社会挑战，还会造成伤害。

为应对这些数据质量挑战，聊天机器人展现出作为可扩展访谈者的巨大潜力。与纸质问卷和面对面访谈相比，它们能够提供互动式、对话式的体验，更有利于收集参与者的数据，尤其是在涉及敏感话题时 [34–36, 54, 72]。研究发现也进一步支持了这一点：由于感到不会受到评判、社交抑制较少且保密性有保障，人们往往会向计算机辅助界面披露更多敏感信息 [10, 53]。然而，这种优势也伴随着过度分享的风险 [74]：受访者可能披露超出问题所需的敏感信息或可识别信息，或者比不使用聊天机器人时自然分享的内容更多。大语言模型驱动的聊天机器人表现得越像人，这一问题就越严重，因为这可能会促使人们披露更多信息 [75, 76]。一个备受关注的行业案例是 Anthropic Interviewer，这是一款人工智能驱动的工具，能够以高质量大规模开展访谈 [22]。然而，聊天机器人引出的丰富细节也加剧了隐私泄露的脆弱性，因为门槛降低，网络搜索大语言模型智能体也更容易重新识别受访者 [33]。此类不必要的信息披露可能使用户面临隐私侵犯或身份盗窃的风险，尤其是在发生数据泄露事件，或数据在未采取充分隐私保护措施的情况下被发布时。

应对过度分享及其引发的隐私风险，常见做法包括在收集数据前取得知情同意，确保采用安全存储流程并向参与者说明，以降低泄露风险、建立信任；数据上传后，则严格实施访问控制和数据最小化 [74]。这些方法通过促进研究过程的透明度和问责制，提供了必要的保障。然而，鉴于研究人员与研究参与者之间固有的权力不对称，这些做法仍不足以全面保护用户，避免其在不情愿或冲动之下披露信息。知情同意的有效性长期以来一直受到质疑 [23]，而参与者在读写能力以及对研究信息的理解方面存在差异，更使这一问题复杂化 [26]。此外，知情同意中需要提供多少细节才算充分，仍存在争议 [63]。

因此，有人提出赋予用户更多控制权，例如控制要分享哪些信息 [4] 以及如何使用所分享的信息 [74]，以此作为修复这种权力不对称的一种方式。然而，这也引发了研究人员对参与者回答真实性的担忧 [74]。目前，关于这一主题的研究十分有限，这仍然是实现伦理研究实践与数据效度之间平衡的一个重要空白。在本研究中，我们旨在通过开展首项考察基于聊天机器人的访谈中隐私控制的研究来弥合这一空白，并围绕两个研究问题展开：（1）参与者将如何在基于聊天机器人的访谈中使用隐私控制；（2）隐私控制将如何影响基于聊天机器人的访谈中的参与者隐私、数据质量和用户参与度？

为此，我们开发了一种访谈聊天机器人系统，提供两种访谈后的隐私控制模式：自由编辑和AI辅助编辑。在这两种模式下，参与者都可以在访谈结束后自由修改自己的回答，并会提前获知这一机会。AI辅助编辑模式的设计动机，源于可用隐私研究文献中反复提及的主题：隐私意识有限，因此需要提供支架来支持理性的隐私决策 [26, 76, 78]。我们的AI辅助模式会扫描整份访谈记录，查找个人身份信息（PII），并提出建议，将个人身份信息替换为类型占位符（例如，[姓名1]），或将其概括为更一般的表述（例如，将“18岁”改为“年轻成年人”）；参与者可以选择接受并应用这些建议。重要的是，我们希望将人工智能作为温和的提醒，弥补隐私意识和隐私读写能力方面的不足，而非将其用作强制性或过度指令性的干预手段。通过这种方式，我们的系统旨在维护用户自主权、恢复用户控制权，并减少权力不对称。

为了实证评估源于这些隐私控制的行为差异，我们开展了一项被试间研究（$N = 188$），包含三种条件：自由编辑、AI辅助编辑和不编辑（对照组）。我们选择设计这一结构化的、基于聊天机器人的访谈，将<目标为="2">在求职面试中使用人工智能作为重点，原因在于该主题具有当代相关性、固有的敏感性 [57, 77]，以及能够接触到一批符合条件的参与者。需要注意的是，在本研究中，我们的分析重点是评估隐私控制方法，而非访谈主题本身。我们收集了编辑前后的对话日志数据，以分析隐私控制的影响，具体考察其对隐私保护、访谈数据质量和参与者参与度的影响。

我们的研究表明，事后隐私控制可以将过度分享转化为小幅且有针对性的编辑，同时不削弱参与者表达内容的实质。AI辅助隐私控制显著提升了隐私保护水平（以个人身份信息的减少量衡量），通过引导用户有针对性地替换或抽象化可识别细节（例如姓名、时间、所属机构、地点），而非直接删除内容或改写得面目全非。隐私保护的提升并未显著损害访谈的实用性：回答仍同样切题、清晰，参与度保持稳定，编辑工作量也较轻且有针对性。相比之下，单靠自由编辑有时会导致内容扩展并披露新的信息，而AI建议则让风险清晰可见，也让安全操作变得简单。我们还观察到认知与行为之间存在差距：尽管不同条件下的参与者对隐私控制和披露信心的感受相似，AI辅助用户却实施了更多保护性编辑，这凸显了醒目、可撤销的文本片段级提示的价值。通过突出透明度、维护用户能动性，并将隐私保护措施融入数据收集流程，我们的研究为应对敏感话题主观数据收集长期面临的挑战，提供了一种实用且符合伦理的

针对敏感话题主观数据收集中长期存在的挑战的对齐回应。

总而言之，我们的贡献如下：

- 我们设计并实现了一种隐私感知型聊天机器人界面，将结构化提问与灵活的会话方式相结合，并集成了两项隐私控制功能，为引导参与者披露敏感信息提供了一种新方法，同时保留参与者的自主权和隐私。

- 我们开展了一项受控实验，以评估提出的两项事后隐私控制功能，并通过对对话日志进行比较分析，评估它们对隐私保护、数据质量和参与者投入度的影响。

- 我们提出了将隐私保护机制整合到基于聊天机器人的访谈研究中的设计指南，使人机交互和社会研究的不同领域能够以符合伦理的方式开展可扩展的数据收集。

2 相关研究

2.1 敏感话题研究中的数据收集挑战

既有研究表明，获取有关敏感话题（包括健康、金融和伦理困境）的准确自我报告颇具挑战。由于社会赞许偏差，参与者往往会少报或歪曲自己的回答 [7, 8, 52, 67], ，而隐私顾虑也会进一步阻碍信息披露 [2]。在线社群和数字平台中也观察到了类似模式，在这些环境中，披露会受到情境规范和感知监视的影响 [39, 40, 65]。此外，社交媒体研究凸显了隐私决策的脆弱性，这往往源于对信息披露的后悔 [71] 以及对受众规模的低估 [38]。近期研究还从认知风格和自动化系统的角度对此进行了探讨，发现风险收益评估会驱动信息共享，而用户友好的界面则有助于改善隐私行为 [5]。

尽管用户表达了强烈的隐私顾虑，但当他们信任接收者时，仍经常披露个人信息，这揭示了态度与行为之间的脱节 [5]。这种现象被称为隐私悖论，指的是个人的行为与其表达的顾虑并不一致，而人们往往以默认设置或感知到的无懈可击为由为此辩解 [46]。研究者通过考察认知和心理影响因素、技术因素以及社会文化因素等，揭示了这一悖论，并强调在数字营销和电子商务的情境下，感知收益超过风险所发挥的作用 [42]。这些问题凸显了自我报告数据收集与印象管理及隐私顾虑之间的交织关系，因此促使人们开发能够减少评判感并建立信任的替代方法。

因此，人机交互研究日益关注如何设计数据收集方法，以应对披露犹豫，并在印象管理与隐私保护之间取得平衡，尤其是在涉及污名化主题时 [3, 31]。本研究通过强调披露后发挥作用的事后隐私机制，推进了这方面的研究，以应对隐私信息过度分享的问题。值得注意的是，我们发现，人们对隐私披露的主观感知与客观检测到的个人身份信息之间存在差异，这呼应了经典的隐私悖论问题；研究还表明，AI辅助编辑能够有效弥合隐私意识方面的差距，并促进更负责任的敏感数据共享行为。

2.2 敏感访谈聊天机器人

研究者已探讨将对话代理用于鼓励用户在敏感领域进行自我披露。先前研究表明，虚拟代理比人工访谈员更能促使受访者敞开心扉，因为人们认为它们不那么爱评判他人，也更容易亲近 [18, 31]。近期研究将这一做法应用于健康与福祉领域，发现聊天机器人有助于人们在心理健康和压力管理方面进行披露 [15, 59]。后续研究考察了聊天机器人交互中的信任、共情与披露，指出其中既有机遇，也存在风险 [44]。针对生成式人工智能聊天机器人在心理健康领域的初步评估显示，用户参与度较高，且能带来积极影响，例如改善人际关系和促进创伤疗愈；与此同时，研究也强调需要增强安全功能并提升类人能力 [61]。此外，学者提醒，聊天机器人介导的同意过程中存在的权力失衡，可能映照出更广泛的自主权问题[55, 73], ，也反映了人们对对话代理期望差距的普遍担忧 [37]。总体而言，这些研究表明，聊天机器人是敏感访谈中前景可观但又存在风险的中介工具，因此必须在披露收益与过度暴露及失去控制的风险之间取得平衡。

我们提出的聊天机器人系统通过集成访谈后隐私泄露检测模块来应对这些问题。该模块能够识别敏感信息，同时不中断会话，从而降低过度披露的风险。手动编辑界面和可选的AI辅助编辑功能，让用户能够掌控自己的披露内容，缓解权力不对等并增强信任，这与先前关于自主权和同意的研究所提出的观点一致。我们通过受控用户研究，将自由编辑和AI辅助编辑与对照组进行比较，实证验证了这些机制如何提升披露质量，并缩小以往聊天机器人研究中发现的期望差距，为敏感访谈情境提供了一种可扩展的解决方案。

2.3 当前隐私同意方法的局限性

静态同意模型因未能提供有意义的用户控制权而受到批评 [6, 62]。动态同意和隐私助推等替代方法展示了如何将持续的情境感知决策融入数字系统 [27, 58]。关于行为广告和位置共享的研究表明，披露偏好高度依赖情境，难以一概而论 [12, 68]。这些设计经常面临在隐私保护与可用性之间取得平衡的挑战，尤其是在实时交互中 [1, 25, 60]。在人工智能和聊天机器人场景中，自动化数据清理技术可以降低数据泄露风险 [9, 43]，但也可能因削弱用户决定哪些数据需要脱敏的自主权而使用户失去权能。近期分析凸显了不断演变的形势：诸如《欧盟人工智能法案》之类的框架要求对高风险系统实施基于风险的治理，并强调透明度和人工监督，以弥补传统同意模型的不足 [24]。因此，既有研究凸显了一个关键空白：如何将自动化与用户主导的

隐私控制以一种有效且尊重自主权的方式相结合。为解决这一问题，应将同意重新构想为贯穿整个交互生命周期的持续过程，并特别关注对话系统中披露的时机。

我们的工作针对这些局限，引入了一种隐私感知聊天机器人，将隐私干预推迟到访谈之后，从而补充主流的预先共享同意模型。该系统的泄漏检测模块以及手动与AI辅助的双重编辑选项，为用户提供灵活的事后隐私控制，在不牺牲自主权的前提下满足情境感知决策的需求。通过在一项关于求职面试中的AI使用的用户研究中评估这些功能，我们旨在证明该方法如何提升透明度与用户能动性，契合《欧盟人工智能法案》对人工监督的强调，并为自动化与用户主导的隐私控制之间长期存在的差距提供切实可行的解决方案。

2.4 聊天机器人系统中面向自主权与隐私的设计

对自主性支持设计日益增长的兴趣源于相关研究，这些研究识别出脚手架支持、选择提供和即时干预等策略，以增强数字自主性 [69]。先前关于知情同意的研究表明，对话界面可以提升用户理解并缓解权力失衡 [73]。此外，隐私配置自主性展示了系统如何在保持可扩展性的同时根据个人偏好进行定制 [55]。相关工作强调在隐私感知人工智能系统设计中让人类保持在回路中 [41, 70] ，并探讨人工智能辅助决策中的信任、透明度与自主权 [28, 29]。近期关于对话式聊天机器人中用户隐私顾虑的文献综述采用社会信息学方法，识别出数据处理、同意和伦理影响等主题，为人机交互设计实践提供参考 [21]。借鉴这些见解，我们的工作在聊天机器人交互中纳入事后、可选的隐私提醒，将自主性支持机制扩展到敏感的自我披露上下文。

这种方法与人机交互领域日益重视的隐私设计趋势一致，优先保障用户能动性，说明轻量级干预措施如何在不损害对话系统可用性或效能的情况下增强自主权。我们的系统通过集成可定制的聊天机器人界面和访谈后编辑选项，进一步践行了这些原则，直接回应了人在回路设计的需求。

3 *EDIT*：增强型披露干预工具

我们开发了一款基于 Web 的应用程序，以下简称增强型披露干预工具（*EDIT*），用于对在线参与者开展半结构化多轮访谈，并在此基础上实现了所提出的两项隐私控制功能。图2展示了所提出的扩展，说明了使用*EDIT*时的示例操作及结果。

3.1 基础聊天机器人系统

我们实现了一个聊天机器人，采用多构成要素编排引擎来管理会话流程，作为访谈的基础系统。编排方式如图 3所示。近期有关对话代理的研究表明，通过提示冻结的大语言模型生成编排信号，有助于对话代理在与用户动态互动的过程中决定应采取哪些行动，从而有效实现特定目标 [14, 17, 32]。受这些研究启发，我们同样采用编排器-执行器设计，以确保聊天机器人会话期间的访谈体验流畅且一致。该引擎以循环方式运行，将每条用户消息依次交由一系列专用大语言模型驱动模块处理，生成符合上下文且遵循访谈流程的回复，并在设计好的半结构化访谈中实现后续问题与主问题之间的平滑过渡（详细提示词见附录B）。我们实现了执行器-审计器状态机循环，因为在早期试点原型测试中，单次生成器偶尔会偏离流程并跳过必需的后续问题，往往连一次访谈都无法完成。加入轻量级质量审计和完成度审计后，系统更可靠地遵循流程，在试点测试中实现了 100% 的完成率和主题覆盖率。

3.1.1 编排引擎。编排引擎负责将上一步的审计结果转换为可执行的提示词，并将其直接传递给执行器，以生成下一轮内容。它控制状态机的平稳转换。对于提出另一个追问或继续下一个主问题等不同任务，编排引擎会相应地获取所需的上下文。

3.1.2 执行器模块。主要的对话轮次由执行器模块生成。收到用户的消息后，编排器会为执行器构建提示词，其中包括当前的主问题、对话历史，以及动态生成的允许操作列表_ACTIONS（例如，ASK_FOLLOWUP、NEXT_QUESTION）。执行器的任务是以结构化 JSON 格式生成响应，指定要采取的操作及相应的话语。该模块的系统提示词明确指示其“引导受访者讲述一个具体的个人故事”，同时保持简洁、自然的对话语气。这是因为，高质量的访谈数据要求受访者描述自己的真实经历和感受，而不是泛泛而谈或使用抽象的表述。参与者仍可选择表示自己没有相关经历，并跳过任何问题。

3.1.3 审计层。为确保遵循研究规程，审计层会评估执行器的输出和用户的输入。该审计层由两个各自独立、由大语言模型驱动的审计器组成。

完成度审计： *auditQuestionPSS* 函数是推进访谈的主要机制。用户提供回答后，系统会调用此审计器，以判断当前主问题相关的预定义后续问题是否已得到充分讨论。系统会向其提供对话历史以及必需后续主题的 JSON 列表。其判定结果要么是允许_进入_下一个问题，要么是需要_更多，编排器会据此决定是转到下一个主问题，还是继续提出后续问题。

质量审计： 为维持会话质量，*au-ditQuestionPresence* 函数会对执行器生成的话语进行质量检查。它会验证机器人的消息是否切题、格式正确（例如，恰好包含一个问题），

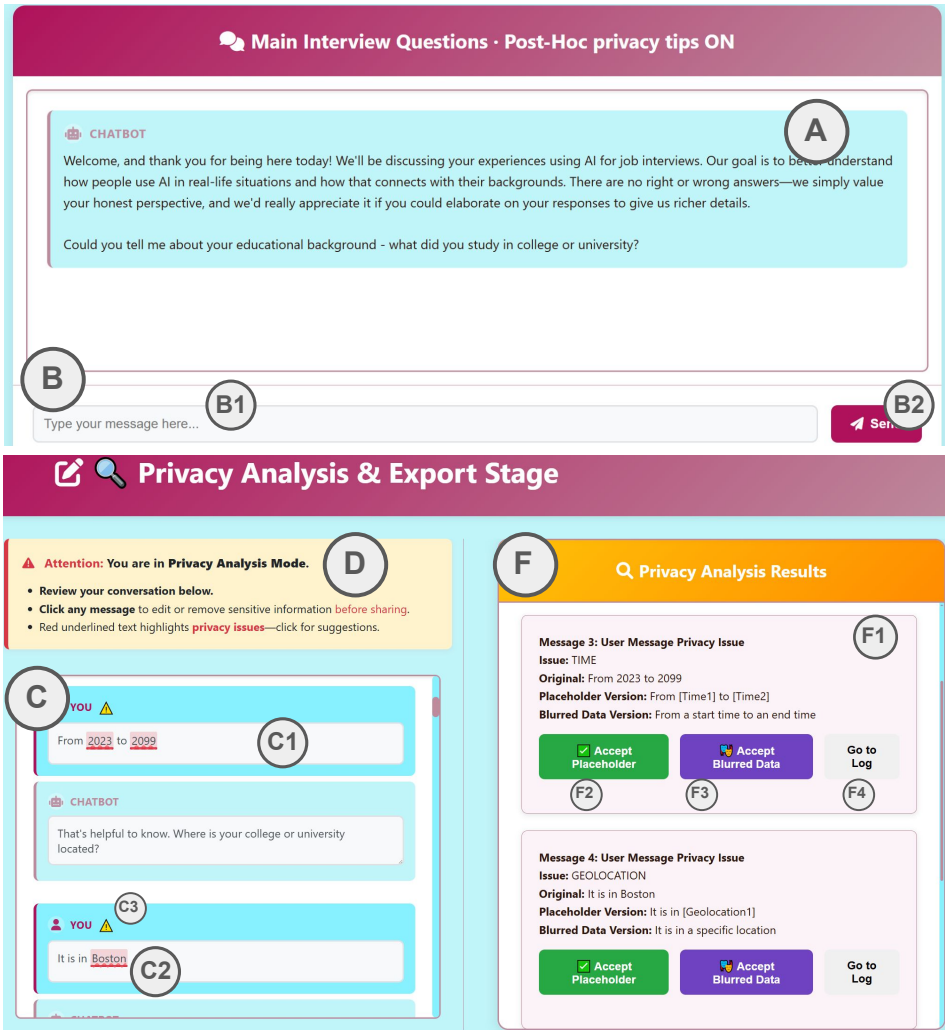

图2：参与者在一次关于*EDIT*求职面试中使用人工智能的访谈期间，体验 AI 辅助编辑的快照。界面起始于聊天机器人界面（上方），其中包含一个可滚动的界面，用于展示用户与访谈聊天机器人之间的聊天记录（A），以及包含文本框（B1）和发送按钮（B2）的文本区域（B）。隐私控制阶段的界面（下方）由一条醒目的提示词（D）组成，用于告知参与者接下来的步骤；一个包含访谈记录（C1）的自由编辑区域（C），支持在文本中突出显示个人身份信息（C2）；以及警示标志中的按钮，参与者点击后即可跳转至相应的隐私问题卡片（C3）。界面通过红色波浪下划线和隐私提示框（F）标记可能包含敏感信息的文本。隐私提示框中包含隐私问题卡片（F1），用于解释风险，并提供三个编辑按钮：将文本替换为占位符（F2）或模糊处理敏感细节（F3）；如果用户想恢复原始日志，还可使用忽略对照组（F4）。在自由编辑组中，参与者只能访问（C）和（D）；而对照组的参与者不会进行访谈后编辑。

且不会过早偏离到其他主题。如果审计未通过，则可以触发重新生成机器人的回复。

3.2 *EDIT* 隐私控制与用户体验设计

除基础聊天机器人系统外，*EDIT* 的核心是可附加到类似访谈聊天机器人系统上的人工智能辅助隐私控制扩展。为了研究不同隐私控制，*EDIT* 被设计为三种不同的实验模式。

自由编辑模式： 访谈结束后，用户会看到一份可全面编辑的完整会话记录。在最终提交之前，他们可以自由添加、删除或修改自己或聊天机器人的任何消息内容。自由编辑模式不提供红色波浪下划线和个人身份信息检测器。初始知情同意书、访谈结束后弹出的窗口，以及

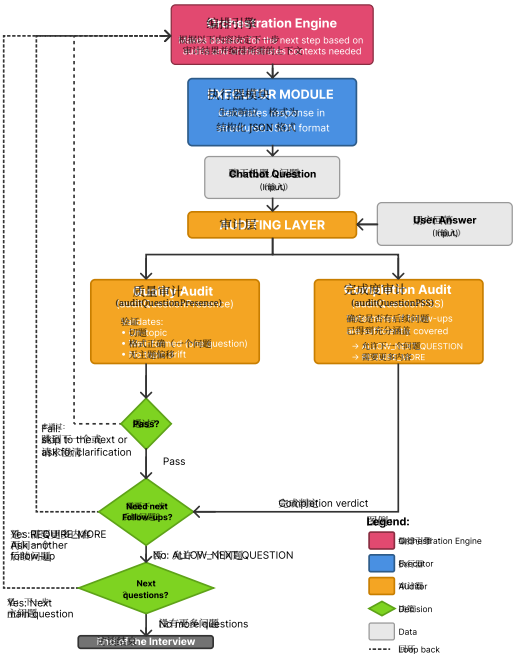

图3：对话式编排可视化示意图。流程开始于编排引擎选择下一个主问题并提示执行器生成结构化JSON响应，并将其作为聊天机器人的问题发送给用户。用户回答后，审计层通过两项LLM驱动的审计结果评估此次交互：质量审计确保响应切题、结构完整且不存在话题漂移；以及完成度审计确定是否已涵盖足够的后续问题。根据这些评定结果，编排器会触发额外的后续问题、请求澄清，或进入下一个主问题。该循环会持续进行，直到没有更多问题，访谈随之结束。

整个访谈过程中聊天机器人界面上的信息性提示词，都会告知参与者可以事后进行编辑。

AI辅助编辑模式：除了自由编辑功能外，此模式还提供由AI辅助的细粒度隐私控制。访谈会话结束后，系统会在对话日志上运行*detectPrivacyWithAI*模块。该模块使用大语言模型，根据 Zhou等人提出的预定义分类体系识别潜在的个人身份信息。 [78]（例如，姓名、所属机构、健康_信息）。完整提示词列于附录B中。随后，系统会向用户展示带有高亮个人身份信息的日志，以及一组选项式建议。对于每个检测到的项目，用户可以选择两种脱敏选项中的一种；为便于界面上的直接说明，这些选项重命名如下：

- **替换（图2中显示的接受占位符）图2：**将敏感文本替换为通用占位符（例如，“我叫约翰”变为“我叫 [姓名1]”。）
- **抽象化（如接受模糊数据 所示，见图2）**：将敏感文本替换为不那么具体但更……

听起来自然的替代表述（例如，“我于2020年毕业”变为“我最近毕业了”）。

与自由编辑模式类似，参与者会在访谈过程开始前、进行中和结束后，获知自己有机会借助人工智能进行事后编辑，并收到明确通知，说明人工智能会进行隐私分析。AI辅助条件提供两种脱敏操作：替换和抽象化；自由编辑条件则仅提供手动文本编辑，不提供下划线标记或人工智能生成的建议。

无编辑模式：该组作为对照条件。完成访谈后，用户直接进入任务后调查，随后提交未经修改的对话日志，期间没有机会查看或编辑。

3.3 实现

EDIT 以客户端-服务器应用程序的形式实现。前端是一个使用 HTML、CSS 和原生JavaScript 构建的静态单页应用程序。后端是一个 Node.js 服务器，使用 Express 框架提供 RESTful API。

3.3.1 后端服务。所有智能模块，包括执行器、审计器和隐私检测器，均通过调用 **OpenAI 模型** [49]，（主要是 **gpt-4。1** [48] 和 **gpt- 4o-mini** [47]）实现。研究提交内容和会话快照的持久化存储由 AWS SDK for JavaScript v3 管理，目标存储位置为启用服务器端加密的 AWS S3 存储桶，并使用 Amazon S3 托管密钥（SSE-S3）。只有经过授权的研究人员才能访问用户数据。

3.3.2 状态管理。服务器端会话状态在内存中进行管理，使用 JavaScript Map对象，其中每个键都是唯一的会话ID。此状态对象存储完整的对话历史、问题完成状态、后续问题覆盖情况及其他元数据。为确保服务器重启情况下的数据持久性，*persistSession-Snapshot* 函数会定期序列化活跃会话的状态，并将其作为JSON文件上传至S3。

4 方法

为了了解不同的隐私干预如何影响用户披露行为，我们开展了一项被试间在线研究（N = 188）。我们考察了用户在三种实验条件下如何与系统互动，以及这些互动如何影响其隐私行为和感知。本研究已获得所在机构机构审查委员会的批准。

4.1 参与者与数据收集

我们通过在线研究平台 Prolific 招募参与者。招募帖将这项研究描述为一次围绕“在求职面试中使用人工智能”展开的会话。符合资格的参与者必须年满 18 岁、英语流利，并通过有关此前在求职面试中使用人工智能经验的筛选问题。完成所有任务的参与者可获得 4 美元报酬，研究持续时间的中位数为 19 分钟。完成速度极快的参与者会被自动筛除，并按照 Prolific 的建议获得筛除级别的报酬。报酬按照 Prolific 建议的时薪标准设定，适用于了

一项持续 15-20 分钟的研究。人口统计信息列于表 5 中附录A。

符合资格后，参与者被随机分配到三个实验条件之一，每种条件被分配到的概率相同。我们共收集到 200 份完整答卷。审阅答卷后，我们剔除了 11 份符合以下任一排除标准的答卷：(1) 对大多数问题仅提供非常简短的回答（例如“是”“否”）；(2) 对大多数问题给出无关回答；(3) 尽管通过了筛选，却表示没有相关经历；或 (4) 提供难以理解的胡言乱语文本。

我们根据试点研究样本中估算的主观评分比较效应量进行了功效分析（n = 107），并确定每种条件下 60 份答卷可提供 80% 的功效，以检测我们关注的效应，显著性水平为 α= 0.05。试点研究采用与主研究相同的研究设计，并对访谈界面进行反复测试和改进。试点样本通过 Prolific 招募，采用与主研究相同的资格标准和研究流程，报酬则与平台建议的标准一致。各条件下的最终分布如下：

- AI辅助编辑模式：60 名参与者
- 自由编辑模式：62 名参与者
- 对照模式：66 名参与者

4.2 研究设计

本研究采用被试间设计，设置了三个条件，分别对应系统的三种模式：自由编辑模式、AI辅助编辑模式和对照模式。每种模式都提供不同的访谈后体验，以测试各种隐私控制机制。我们选择了三条件被试间设计，所有隐私控制均在访谈结束后实施，而非访谈过程中。我们在初始知情同意书中告知参与者访谈后可以编辑，并在访谈结束后立即通过清晰的弹出提示再次说明，使这一过渡自然且提示明确。这样可以避免访谈过程中出现实时警告或提示，从而让我们能够以贴近自然情境的方式观察参与者的披露行为；同时，访谈后的控制措施使参与者能够自主修改转录文本，并让我们得以分离各隐私控制条件对研究问题的因果效应。此外，收集编辑前和编辑后的访谈记录，对于研究参与者在获得隐私控制后如何实际改变披露内容是必要的。为尽量降低风险，我们在收集原稿（编辑前）转录文本前，另行征求参与者主动选择同意；即使拒绝收集这些文本，参与者仍可获得全额报酬。通过这种方式，研究设计既能获取回答研究问题所需的数据，又能避免更具侵入性的做法（例如强制收集编辑前的转录文本，或采用默认收集此类数据的退出式同意）。因此，我们认为这是一种侵入性相对较低的研究方法。

访谈和特定模式的任务完成后，每位参与者都会填写一份任务后调查，以收集其主观反馈。调查包含两组问题（完整的访谈脚本、筛选题目和任务后调查问题见附录A）：

- **所有模式的共同问题**：所有参与者都回答了一组关于整体体验的问题。这些问题衡量了他们普遍的隐私顾虑、犹豫的时刻、对数据收集的感受，以及与聊天机器人分享个人信息时的整体舒适度。

- **针对编辑模式的问题：**为评估特定隐私干预措施的影响，自由编辑组和 AI辅助编辑组回答了一组根据其体验量身定制的附加问题。这些问题用于衡量参与者如何看待编辑功能对其隐私控制感、对隐私保护的理解、编辑后分享信息的信心，以及其编辑行为和理由的影响。

4.3 研究流程

每位参与者都要经历四个主要阶段。图 4概述了整个流程。

引导与知情同意：参与者首先进入介绍页面，页面概述了这项研究，并提供了机构审查委员会批准文件的链接。随后，他们进入知情同意页面，必须同意相关条款才能继续。

资格筛选：参与者回答了一组简短的筛选问题，以确认其是否符合研究资格。为避免偏见，我们以中立的语气呈现问题，并要求参与者从四个选项中选择最符合自身经历的一项，而不透露哪些选项是预期答案。报告称曾在求职面试中使用人工智能、曾使用人工智能准备求职面试或两者兼有的参与者通过筛选。其余参与者均未通过筛选，系统会自动将其重定向至研究结束页面，并根据平台建议向其支付 0.14 美元的筛选补偿。我们仅招募了美国境内的参与者。

对话式访谈：符合条件的参与者进入主聊天界面，开始半结构化访谈。在访谈中，聊天机器人会依次提问，并提出后续问题。AI代理引导参与者回答一组预先设定的问题，了解他们在求职过程中使用人工智能的经历。这些问题分为六个固定的问题组：（1）第1组：教育背景；（2）第2组：当前工作和招聘流程；（3）第3组：面试中的AI使用；（4）第4组：使用AI时的紧张经历；（5）第5组：AI使用的适当性；（6）第6组：隐性AI使用。这些问题组涵盖了从个人背景到求职面试中人工智能使用相关话题的信息，话题范围从一般到敏感，参与者的回答中也更可能包含个人身份信息。系统的编排引擎负责管理对话流程，并提出必答的后续问题，以确保在进入下一个主问题之前，充分探讨每个话题。我们禁用了复制粘贴功能，以减少参与者使用人工智能生成回答的情况。完整访谈方案见附录A。

访谈后任务与问卷：在完成最后一道访谈问题后，参与者参与了以下其中一种模式的任务。最后，被分配到两种隐私控制条件的参与者被请求额外同意分享其原稿日志和编辑后的日志，以促进研究

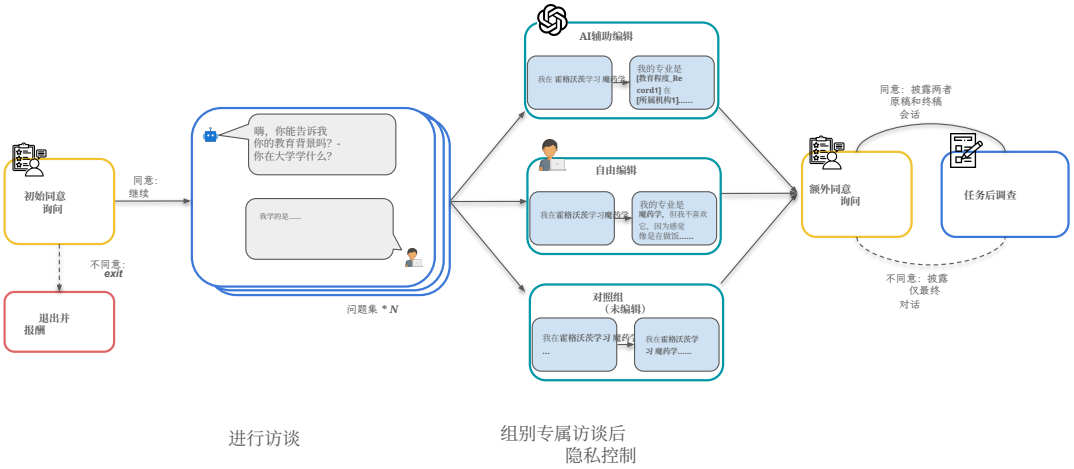

图 4：研究流程的主要阶段和条件：1）向参与者展示半结构化聊天机器人访谈的初始知情同意书；2）阅读知情同意书后，参与者被随机分配到三个条件之一，以相同概率进行访谈；3）访谈结束后，根据所分配的条件，参与者会获得三种访谈后隐私控制之一：AI辅助编辑、自由编辑或不编辑（对照组）；4）参与者完成额外的同意相关询问；5）所有参与者最后完成任务后调查

分析，并保证保密性。他们可以选择接受分享两份日志，或拒绝而仅分享编辑后的日志。所有参与者均同意分享两份日志。

- **自由编辑模式：** 参与者看到的是一份可完全编辑的对话转录文本，并被要求在提交前按自己的意愿进行修改。
- **AI辅助编辑模式：** 参与者看到自己的转录文本，以及一份 AI 生成的隐私分析，其中标出了潜在的个人身份信息。随后，他们可以选择保留原文、将其替换为通用占位符，或替换为更抽象、较难识别的版本。参与者也可以像在自由编辑模式中一样自由编辑。
- **对照模式：**参与者直接进入最终问卷。

4.4 数据分析

除了个人身份信息披露的数值结果外，我们还考察了人工智能辅助编辑组和自由编辑组参与者的编辑行为，并分析其编辑模式是否存在差异。此外，通过引入 Xiao 等人提出的三维评分标准 回答质量指数（*RQI*） [74]，该标准将相关性、清晰度和具体性视为整体回答质量的组成部分，我们从三个方面量化数据质量，以比较不同隐私控制对收集的数据的影响。我们将在第 4.4.2 节中详细说明 RQI。

4.4.1 编辑模式。 在分析参与者多样的编辑行为并找出共同模式之前，将编辑行为编码为特定类别，例如 删除内容，可以减少编辑中的个体差异和噪声，并按群体揭示其背后的动机。编辑操作的编码手册见 表1。

为尽量减少参与者行为编码中的偏差，我们采用了迭代式开放编码流程。首先，我们分层抽取了 50 条回答（在不同实验模式和问题组之间均衡分布）。两位作者分别对这组回答进行开放式编码并制定初步编码手册，随后共同讨论分歧、合并重叠编码，并完善类别定义和判定规则。使用修订后的编码手册，两位作者又分别对另一组 50 条回答进行编码。采用 Cohen’s κ 计算的评审者间信度表明，$\kappa = 1.00$在第二组样本中达成了完全一致。之后，一位作者将最终版编码手册应用于剩余语料库。

4.4.2 回答质量指数（*RQI*）。我们采用 Xiao 等人 [74] 提出的三维评分标准来操作化答案质量：相关性（0–2）、清晰度（0–2）和具体性（0–2）。单条回复的指数为三项得分的乘积

$$\mathrm{RQI} = \sum \mathrm{Relevance}[i] \times \mathrm{Clarity}[i] \times \mathrm{Specificity}[i],$$

取值范围为 0–8。我们对 message_原稿和 message_终稿均进行了评分。

为了扩大评分标准的一致应用，我们使用大语言模型辅助评分。具体而言，我们采用了**GPT-5 Thinking** (Ope- nAI)，并使用确定性解码（temperature= 0, top_p= 1）。提示词逐字重述了评分标准，要求每个维度使用{0,1,2}的整数标签，禁止使用外部知识，并要求以结构化的 JSON 输出提供简短（少于 30 个词）

针对原稿和终稿的回答，提供基于证据的简要理由。大语言模型首先在不知晓作者标签的情况下为校准集标注；随后，我们在一个分层子集上将大语言模型的标签与人工金标准进行比较，并反复调整提示示例，直至信度趋于稳定。在留出的校准集上，人类与大语言模型的一致性达到较高 水平，Cohen's $\kappa = 0.652$。

5 结果

本节介绍我们针对所提出的隐私控制措施得出的结果。我们首先评估 AI 辅助干预在隐私保护方面的整体有效性，随后深入探讨自由编辑与 AI 辅助编辑之间不同的编辑模式，以及 AI 辅助对用户的异质性影响；接着，我们测量并比较了所有条件下的数据质量和用户参与度，结果表明差异并不显著。我们采用组间 Welch t 检验。对于方差分析，自由度为 $F(2, 185)$。

5.1 隐私干预对个人身份信息披露的有效性

我们发现，在终稿访谈聊天记录中，AI 辅助编辑组每条消息的平均个人身份信息数量最低（$M = 0.107$），低于自由编辑组（$M = 0.259$）和对照组（$M = 0.247$）；方差分析结果为（$F = 30.747$, $p < .001$, $\eta^2 = 0.249$）。

我们发现，与**自由编辑**组相比，**AI 辅助编辑**条件下每条消息的平均个人身份信息减少率显著更高（t 检验：$t = -10.570$, $p < .001$）。

AI辅助编辑组参与者在整个语料库中的个人身份信息净减少率达到 **41.2%**，其中 41/60（68.33%）的参与者至少使用过一次 AI辅助功能，涉及时间、所属机构、地理位置、教育记录、姓名和健康信息等个人身份信息类别。然而，自由编辑组的减少率为 **-2.6%**（即净增加），如图 5所示。在参与者层面也呈现出同样的趋势：AI辅助编辑组中有 71.2%的参与者主动减少了其个人身份信息，而自由编辑组中这一比例为 22.6%。

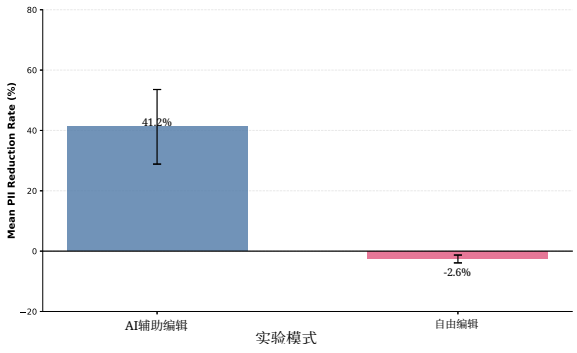

图 5：不同实验条件下个人身份信息减少率的均值。AI辅助编辑条件下，个人身份信息显著净减少；而自由编辑条件下，编辑后个人身份信息略有增加。误差线表示标准差。

5.2 不同个人身份信息披露基线用户的隐私保护效果

我们考察了哪些用户从 AI辅助干预中获益最多。在 AI辅助编辑组中，我们发现，基线个人身份信息暴露程度较高的参与者实现了更大的绝对减少量（原稿个人身份信息与总减少量：皮尔逊相关系数 ≈ +0.15）。然而，这些参与者实现的比例减少幅度反而较小（原稿个人身份信息与减少率：r ≈ -0.21，对于基线值非零的参与者），如图6所示。

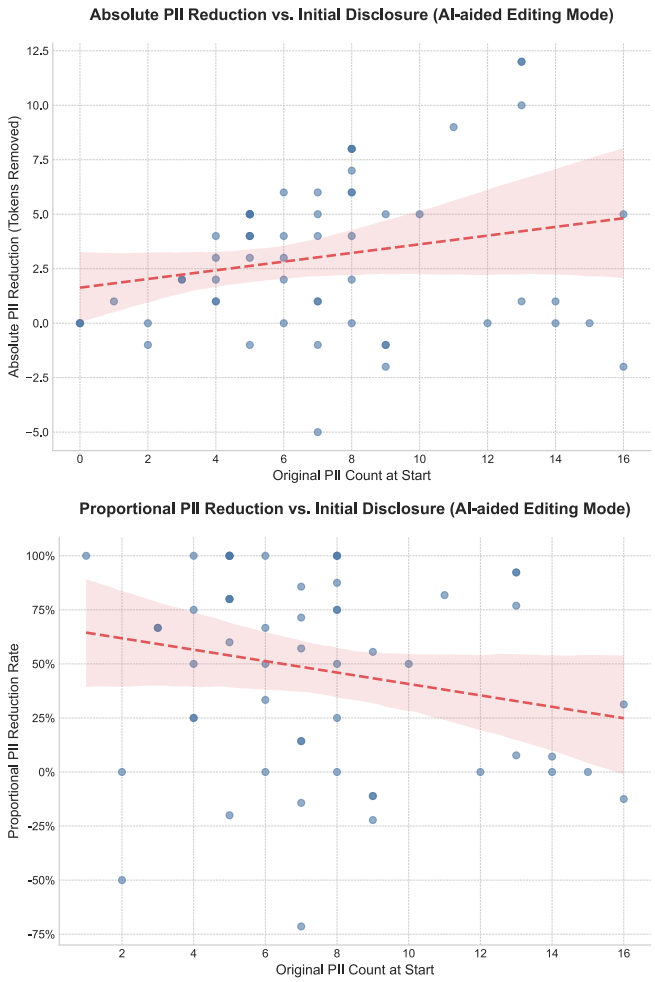

图6：参与者初始个人身份信息数量与AI 辅助干预效果之间的关系。（上）初始个人身份信息较多的用户，绝对移除的词元数量也更多。（下）初始个人身份信息较多的用户，移除的个人身份信息占其个人身份信息总量的比例较小。

四分位数分析凸显了这一趋势。基线处于最低四分位数的参与者平均从 3.52 个个人身份信息标记开始，平均减少率为 62%。相比之下，处于最高四分位数的参与者平均从 13.33 个个人身份信息标记开始，尽管删除的

标记绝对数量更多，但减少率仅为 36%。这种差异凸显了个人在隐私需求和风险感知方面的不同。

5.3 编辑模式

手动编辑的性质在 AI辅助编辑和自由编辑条件之间存在显著差异，这表明无指导的编辑可能适得其反。在 AI辅助编辑组中，用户根据 AI建议进行编辑，其中 95.48% 的手动编辑导致个人身份信息减少。相比之下，自由编辑组中只有 19.23% 的编辑减少了个人身份信息，11.54% 的编辑反而增加了个人身份信息，而 69.23% 的编辑对个人身份信息没有影响（见 图7）。

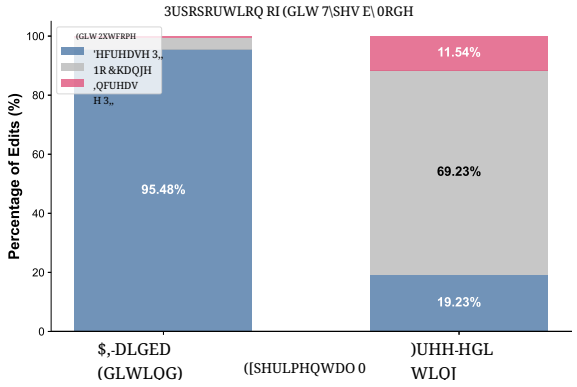

图7：编辑结果的比例分布。AI辅助编辑绝大多数减少了个人身份信息，而无指导的编辑大多保持中立，或在某些情况下增加了个人身份信息。结果经过四舍五入

表1 总结了所有经过编辑的回答中出现的各类编辑操作。表2 展示了 AI辅助编辑与自由编辑会产生不同的编辑模式。AI辅助组中的所有编辑操作都集中在“替换为类型占位符”和“从抽象到一般信息”这两类操作上（见 表2），分别对应系统提供的两种 AI建议。值得注意的是，我们观察到少数情况下，参与者会先采纳文本抽象化的 AI建议（例如，将“亨德森学区”抽象为“学区”；Henderson 是用于匿名化用户数据的虚构标识符），然后再手动编辑回答以补充更多细节（例如，添加“当地”一词，将其改为“当地学区”）。这展现了一个有趣的人机协作案例：人类会主动提高回答的具体性。

相反，自由编辑组的编辑操作集中在另外五个类别上（表2）。其中两个与隐私保护有关，分别是“删除内容”和“替换为不可识别的占位符”，体现了一种非此即彼的隐私保护方式。其余三个类别——“修正格式/错别字”“添加新内容”和“更改答案”——与隐私无关，在某些情况下甚至可能增加信息披露。

表1：编辑操作编码手册。可识别信息（如公司名称）已替换为与原始格式相似的虚构示例。

Code	代码说明	示例
替换为类型占位符	用户替换了个人信息为该信息输入为占位符	[{‘原稿’: ‘2023年’, ‘终稿’: ‘[时间2]’}]
抽象为一般信息	用户将个人信息抽象成更笼统的版本，减少了细节	[{‘原稿’: ‘林肯学校地区.’, ‘终稿’: ‘当地学校学区.’}]
移除内容	用户完全删除了某些内容	[{‘原稿’: ‘我在其中一所大学里做到了在美国..’ , ‘终稿’: ‘ ’ }]
添加新内容	用户添加了新内容使用编辑功能来提供更多信息，进行澄清和阐述，而不删除现有内容。	[{‘原稿’: ‘’, ‘终稿’: ‘所以我确实不隐藏我的使用情况因为我想要’}]
编辑为无法识别的占位符	用户替换了个人无法识别的信息作为占位符的文本	[{‘原稿’: ‘Redwood Corporation’, ‘终稿’: ‘*****’}]
Fix 格式/拼写错误	用户未添加新内容使用编辑功能，但修改了文本以纠正语法错误或标点符号。	[{‘原稿’: ‘大量练习’, ‘终稿’: ‘大量练习’}]
更改答案	用户更改了答案而不是删除已提供的答案或添加新的内容	[{‘原稿’: ‘2018年’, ‘终稿’: ‘2023年’}, {‘原稿’: ‘2018年.’, ‘终稿’: ‘2023年.’}]

表2：已编辑用户消息中编辑代码按条件的分布。百分比为条件内占比。

Code	AI辅助编辑	自由编辑
从抽象到一般信息	35 (23.6%)	2 (8.3%)
添加新内容	0 (0.0%)	7 (29.2%)
更改答案	0 (0.0%)	3 (12.5%)
修正格式/拼写错误	0 (0.0%)	7 (29.2%)
替换为类型占位符	111 (75.0%)	0 (0.0%)
替换为不可识别的占位符	0 (0.0%)	4 (16.7%)
删除内容	2 (1.4%)	1 (4.2%)

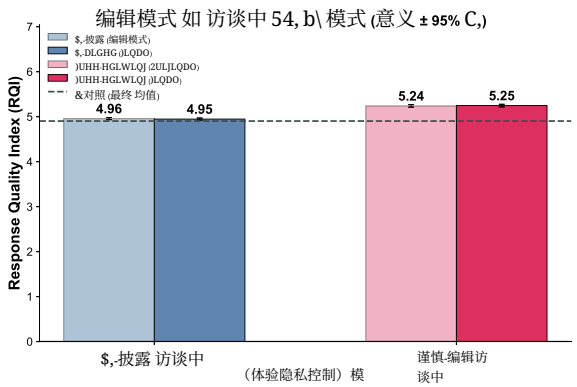

图 8：按模式划分的总体 RQI（均值 ± 95% 置信区间）。AI辅助编辑和自由编辑对数据质量的影响均较小，且在 RQI 衡量的数据质量方面没有显著差异。

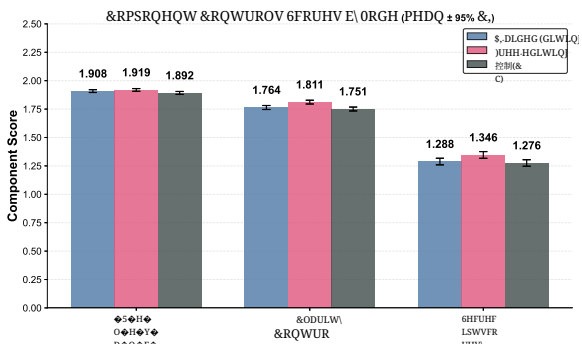

图9：按模式划分的总体构成要素得分（相关性、清晰度、具体性）（均值 ± 95% CI）。在这两个维度上，三种模式之间均未观察到显著差异。

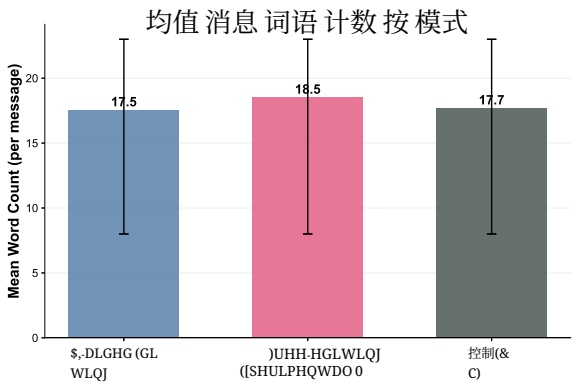

图10：按模式划分的消息词数（均值，误差线表示第25至第75百分位数）。三种模式之间的用户参与度未观察到显著差异。

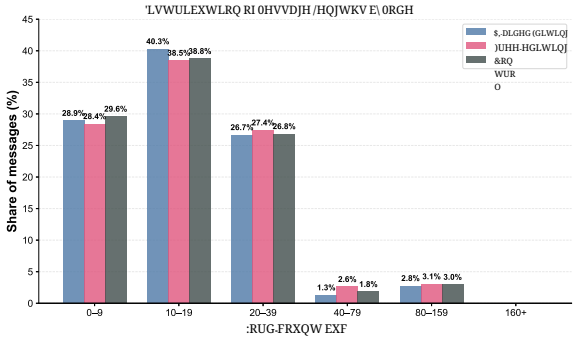

图11：按模式划分的消息长度分布（各区间内消息所占比例）。长尾分布表明，用户倾向于用一个简短句子作答。超过94%的消息长度在40个词以内，且未观察到特别长的回答（≥ 160）。

5.4 回答质量

我们在参与者层面评估回答质量，采用综合的**回答质量指数（RQI）**[74] 及其格赖斯会话准则组成部分 [19] (相关性、清晰度、具体性）。每位参与者的 RQI 为其所有单条消息 RQI 的平均值。三个条件之间的 RQI 未观察到显著差异（**自由编辑**：M = 5.24，95%置信区间 [5. 06，5。42]，n = 62；**AI辅助编辑**：M = 4 95, 95 [4 80, 5 10] n = 60 = 4 92, .% CI ..,; 95 [4 77, 5 08] n = 66 8 **对照组**：M . % CI . . , ）。这一趋势见图。相关性、清晰度和具体性的趋势见图9。

5.5 用户参与度

词数我们使用消息级词数作为参与度的代理指标。词数区间便于比较不同模式，并确保文本长度范围相近。四分位距（Interquartile Range，IQR）是第一四分位数末端与最后四分位数起点之间的距离，用作衡量统计离散程度的指标；它忽略极端值，聚焦于数据集的核心部分。不同条件下的分布大致相似，中位数为**14**词，IQR为**8–23** 词（表3）。**自由编辑**产生的平均消息略长（M = 18.5，SD = 16.38），高于**AI辅助编辑**（M = 17. 5，SD = 15.32）和**对照组**（M = 17.7，SD = 15.94），即分别多出 +1. 0词和 +0.8词（图10、表3）。如图11所示，分桶分布显示，自由编辑中极短消息（0–9词）所占比例（28.4%）略低于AI辅助编辑（28.9%）和对照组（29.6%），而20–39词（27.4%对比26.7–26.8%）及40–79词区间（2.6%对比1.3–1.8%）的比例相应增加。极长消息（≥80 词）在所有模式中仍很少见（≈2.8–3.1%）。

表3：按模式划分的消息级词数：分布摘要。

Mode	Mean	SD	Min	P25	中位数	第75百分位数	第90百分位数	Max
AI辅助编辑	17.5	15.3	1	8	14	23	34	145
对照组	17.7	15.9	1	8	14	23	34	144
自由编辑	18.5	16.4	1	8	14	23	34	113

编辑率。我们将编辑活动量化为每个问题组中被编辑的用户消息所占的比例。如图12所示，

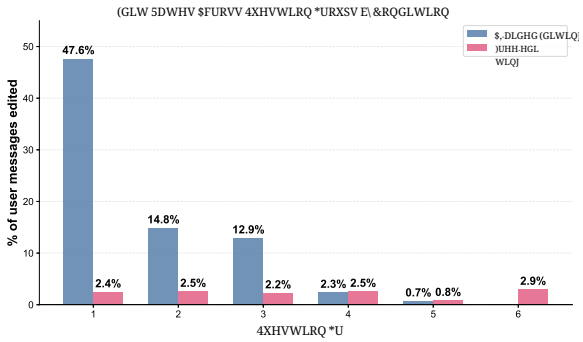

图12：不同问题组中各模式的编辑率分布。与个人背景更相关的前三个问题组中，AI辅助编辑的编辑率明显更高。

AI辅助编辑将编辑集中在访谈早期——第1组：**47.6%**，第2组：**14.8%**，第3组：**12.9%**——此后逐渐降至接近零（≤**2.3%**）。相比之下，**自由编辑**在各组中始终处于较低水平（≈**2–3%**：**2.4%**、**2.5%**、**2.2%**、**2.5%**、**0.8%**、**2.9%**）。我们注意到，前三个问题组聚焦于参与者的背景和经历，因此涉及更多个人身份信息，例如姓名、学校和所属机构。相比之下，最后三个问题组聚焦于参与者的态度，因此包含较少的个人标识符。两种条件下呈现出的不同模式表明，AI辅助编辑有助于用户将编辑集中在存在隐私风险的部分，而不是进行分散编辑；后者可能无法降低隐私风险，还可能影响数据质量。

5.6 主观感知与意愿（李克特量表）

我们考察了参与者自我报告的披露舒适度和讨论敏感话题的意愿，也评估了他们对隐私控制的感知以及对功能的理解。对于Q3（舒适度），对三种模式进行的单因素方差分析未发现显著差异（$F = 1.3494$，$p = 0.262$，$\eta^2 = 0.0145$）。对于 Q4（未来的敏感话题），方差分析同样未显示显著差异（$F = 1.6023$，$p = 0.204$ $\eta^2 = 0.0170$ ，）。

随后，我们比较了两种编辑方式中语义匹配的条目：Q7/Q8/Q9（**自由编辑**）与 Q12/Q13/Q14（**AI辅助编辑**），这些条目均采用五点李克特量表评分，分数越高越好。描述性统计表明，两种方式都易于理解，并提升了感知控制感，均值约为 4。根据表4，这些配对问题之间没有统计学上的显著差异。这些主观感受与行为发现相辅相成：参与者认为，在两种编辑条件下，他们都能自在地管理隐私；与此同时，AI辅助模式产生的残留个人身份信息标记显著更少。

6 讨论

我们的研究结果提供了有希望的证据，表明 AI 辅助的访谈后编辑功能改善了隐私保护，表现为个人身份信息显著减少，同时没有显著影响数据质量或用户参与度。本节将讨论这些研究结果对设计符合伦理的访谈聊天机器人的启示，同时兼顾获取高质量数据以确保研究效度的需求。

6.1 我在披露吗？感知安全性与客观个人身份信息编辑之间的差异

参与者的感知舒适度和信心并不一定与他们的客观隐私编辑情况相符。尽管自由编辑组和 AI 辅助组的用户报告了相似的舒适度（根据任务后的反思），但他们最终的访谈记录却存在差异：AI 辅助组的回复中残留个人身份信息标记远少于自由编辑组和对照组（第 5.1 节）。参与者在得到明确提醒后减少了更多个人身份信息，这表明对照组和自由编辑组额外分享个人身份信息，可能源于缺乏意识。参与者在 AI 辅助条件下减少了更多个人身份信息，这表明对照组和自由编辑组中的额外披露，与其说是主动承担风险，不如说是默认选项的影响所致。在任何干预之前，隐含的默认做法是保持访谈记录不变，而关于隐私冷漠的既有研究[11] 表明，即使提供手动编辑选项，许多人也会顺势维持现状。AI 建议有效降低了偏离这一默认做法所需的努力：参与者只需付出极少的认知成本和交互成本，就能接受有针对性的隐私保护编辑，从而“解锁”更多个人身份信息的删减，而无需相应提高对隐私的关注。换言之，用户在无意识中分享了超出充分知情情况下原本意愿分享的个人信息，这凸显了参与者与研究人员之间的权力不对称。

我们将 AI辅助模式下有效的隐私保护归因于两种相互强化的机制。第一种机制涉及增强的意识：AI建议会凸显出一些具体文本片段；这些片段单独披露时可能看似无害，但在审阅过程中结合起来考虑时却可能显得具有风险（例如姓名 +地点 + 日期），从而促使参与者删除或抽象化这些信息，而这些内容往往会被手动检查遗漏。第二种机制与便利性有关：通过降低隐私处理的认知负担，AI建议使信息清理变得如同单击一下（替换/抽象化）般简单，从而让参与者能够进行更有针对性、影响更大的编辑，而不是进行宽泛且不精确的删除。

相比之下，我们观察到，尽管自由编辑条件下的参与者报告的隐私控制感和分享信心，与 AI辅助条件下的参与者同样积极（表4），但他们的编辑往往只是表面上处理了隐私保护问题，有时甚至因提供更详细的回答而引入了额外的个人身份信息（第5.3节）。总体而言，我们认为自由编辑模式可能会助长一种虚假的安全感。这种感知与行为之间的差距表明，隐私风险并不是一种可靠的直觉感知；可用控制措施必须将模糊的安全直觉转化为文本片段级具体且可执行的编辑。

6.2 平衡隐私、参与度与数据质量

一个普遍的迷思认为，为用户提供有意义的隐私控制会带来降低数据质量的风险，这一假设体现在研究实践 [74]以及公司使用暗黑模式来使

表4：人工智能与自由编辑（匹配条目）。独立组；无被试内对比。三种模式在这些主观模式上没有显著差异。

配对 / 条件	计数	Mean	中位数	SD	t	p
Q7: 编辑功能提升了我的隐私感对照组。（自由编辑）	62	3.92	4.00	0.88	−1.1252	0.2628
Q12: AI辅助编辑提升了我的隐私控制感。（AI辅助）	60	4.08	4.00	0.91		
Q8: 我清楚地了解如何使用编辑功能来保护我的隐私。（自由编辑）	62	4.11	4.00	0.79	−0.1187	0.9057
Q13:我清楚地了解了AI建议如何帮助我保护了自己的隐私。（AI辅助）	60	4.12	4.00	0.90		
Q9: </style'>自由编辑让我对自己分享的内容更有信心。（自由编辑）	62	4.00	4.00	0.87	0.0965	0.9233
Q14: </style'>AI辅助编辑让我对自己分享的内容更有信心。（AI辅助）	60	3.97	4.00	1.06		

退出变得不便 [76]。我们的研究结果挑战了这一迷思，表明为用户提供用于清理可识别个人信息的定向指导，能显著提升隐私保护，且不会损害数据质量（第5.4节）或用户参与度（第5.5节）。换言之，参与者仍继续提供经过思考的详细回答，并未删除关键信息或操纵其答案。

除了总体趋势之外，对编辑模式的详细分析进一步揭示了两种隐私控制机制对数据质量的影响。在自由编辑条件下，参与者常将此功能用作通用修改工具，频繁添加更多细节（“添加新内容”编码）。例如，一位参与者最初误解了某个问题，后来在访谈后自由编辑时补充了缺失的答案。然而，当编辑涉及隐私时，自由编辑模式会以不同方式损害数据质量。在“更改答案”模式中，参与者修改答案，因而引入了可能不实的信息。在“将内容删改为无法辨认的占位符”模式中，尽管这与AI辅助条件下观察到的“将内容删改为占位符”模式类似，但生成的文本可读性较差。最后，在“删除内容”模式中，参与者直接删除整个句子，没有留下任何痕迹表明信息曾被有意隐藏。原稿与编辑后回答的匿名化示例对见附录E。

与自由编辑相比，AI辅助编辑的一大优势是“从抽象到一般信息”和“涂改为类别占位符”这两种模式更为常见。这种模式在披露细节的程度与保护参与者隐私的需求之间取得了平衡。我们承认，在尽可能删改个人身份信息与保留实验所需的收集数据中的必要信息之间存在张力，因此设计了 AI辅助编辑，建议将内容替换为可识别性较低但仍包含一般信息的片段，以维持所收集用户数据的数据质量。我们的 AI辅助编辑机制在大幅减少个人身份信息的同时，仍能维持 RQI 分数和参与度。

尽管这种情况并不常见，但尤其值得关注的是，我们观察到了一些与仅接受 AI 建议或进行手动编辑截然不同的模式，这表明可能存在一种“双重编辑”路径。这些模式表明，参与者接受了抽象化建议，随后又通过手动编辑补充了更多细节。以下是我们从数据中观察到的一些例子：（1）原始的 AI 抽象化建议将我去了霍格沃茨大学[1] 改写为我去了一所大学，而终稿句子是我去了一所大学 在线；（2）对湾区的原始抽象化表述是一座大城市，之后又被进一步修改为加利福尼亚北部；（3）对哈利·波特教授 的原始抽象化表述是一个人，之后又被进一步修改为一位化学教授。将这些结果与自由编辑的结果相比较，可以看出，这与我们观察到的非此即彼行为形成了对比（保留原文不变或大幅删改）。这表明，如果不直接套用 AI 建议，参与者很可能会将 AI 建议作为提示，并结合自身偏好与 AI 建议来修改答案，形成一种协作式的折中披露方式；当参与者拥有一定程度的自主权时，这种方式既不易识别个人身份，又仍具有分析价值。

总而言之，证据表明，人工智能辅助脚手架能够实现双赢：帮助用户保护隐私，同时让研究人员保留分析所需的上下文信号。

我们的目标并非通过简单粗暴的删除来最大限度地移除个人身份信息，而是借助人工智能提示，提醒参与者注意潜在的个人身份信息，并将自主权交还给他们。当隐私与效用发生冲突时，我们认为应始终优先保障用户隐私，尽管要找到真正平衡的解决方案，可能需要进行一些协商和集体优化。我们的研究迈出了颇有希望的第一步：与自由编辑相比，AI辅助编辑显著减少了个人身份信息，同时保持了 RQI 分数和参与者参与度，这表明在隐私干预措施提供的脚手架支持和指导下，可以将更多信任交托给参与者。不过，在某些情况下，细致且因人而异的信息可能对下游分析十分重要；此时，即使高层次的质量指标保持稳定，过度删改仍可能降低数据丰富度。未来研究应致力于更好地兼顾参与者和研究人员的优先事项，既

[1]合成个人身份信息用于替换参与者提及的实际个人身份信息，仅作说明之用。

维护参与者的自主权及其保护隐私的能力，也通过协作协商找到兼顾隐私与用户数据丰富度的折中方案。我们的结果也指出了改进方向。例如，可以结合 AI辅助和自由编辑条件各自的优势，比如鼓励参与者添加新内容，同时提醒他们在这一过程中警惕额外披露个人身份信息，以避免过度分享。

6.3 探索聊天机器人访谈中隐私控制的设计空间

类人行为 [76] 以及大语言模型的社会常识 [45]使其非常适合用于大规模开展深入访谈，为依赖人类受试者研究的领域提供了一种新颖的方法，尤其适用于涉及敏感话题的研究。然而，先前的研究指出了参与者过度分享的风险，我们的研究结果也印证了这一担忧。我们的工作初步探索了访谈聊天机器人设计中的隐私控制，并取得了积极成果，鼓励未来研究沿这一方向继续深入。我们还概述了更广泛设计空间中的其他机遇，供未来研究探索。

访谈聊天机器人与人工访谈员的混合或替代应用。尽管现有研究证明了聊天机器人在敏感访谈情境中的实用性 [31, 61], , 但人工访谈员仍不可或缺，因为他们能够提供情感支持和必要资源，例如实时共情、语境细微差异，以及访谈员与受访者之间的融洽关系。除了这些情感层面的益处，混合工作流程还会带来额外的隐私控制需求和机遇。例如，聊天机器人可以充当初步倾听者，让参与者按照自己的节奏披露信息，并在任何内容呈现给人工访谈员之前，使用可配置的隐私设置（例如自动屏蔽姓名、地点和可识别实体）。随后，系统可以向人工访谈员提供去标识化摘要或选择性披露的片段，让参与者有机会在分享之前批准、否决或进一步编辑特定内容。聊天机器人可以标记可能敏感的内容片段，并明确询问参与者是否应将这些片段转交给人工访谈员；人工访谈员则可以解释其中的权衡、回答澄清问题，并共同商定适当的细节程度。通过结合双方的优势，在某些情境下采用协作工作流程或替代应用，可以作为传统人工主导访谈的补充，而非取而代之。例如，在人工访谈员安排后续访谈之前，可以将聊天机器人用于预筛查或分诊工具。

控制时机。 除了访谈结束后立即进行编辑的功能外，未来研究还可以改变干预时机，并考察其对隐私保护、数据质量和用户参与度的影响。例如，在参与者与聊天机器人交谈时发送即时提醒，这可能带来即时反馈的好处，也可能产生抑制效应的风险；还可以建立长期渠道，以便在数据收集后进行隐私控制。

隐私保护方法。 隐私保护程度通过对话日志中共享和保留的个人身份信息数量来评估。这种方式能直观地向用户说明隐私风险，也常被用作隐私法规（如《通用数据保护条例》）中的衡量标准。然而，它无法提供匿名化的理论保证 [66]。为弥补这一局限，未来研究应探索其他隐私定义，例如差分隐私（DP），并开发能够提供此类形式化保证的机制。例如，生成差分隐私合成数据，能够在个体层面保护隐私，同时在数据聚合时维持群体层面的实用性。主要挑战在于如何向最终用户传达差分隐私的隐私保护优势 [13] ，以及如何向研究人员说明隐私与实用性之间的权衡，以便他们在充分知情的情况下选择隐私预算参数，例如 ϵ [56]。

持续参与。在我们的研究中，每位参与者仅在特定条件下与聊天机器人完成了一次会话。开始会话时，他们尚未体验访谈后隐私控制功能。我们的研究结果表明，接触这些功能可能会影响参与者的感知安全性和披露信心。不过，这一影响仍需在受控环境中进行正式测试：让参与者在相同条件下反复接受访谈，以便更系统地评估该方法的长期影响，并深入了解参与者披露行为背后的机制。

7 局限性

尽管我们的研究初步表明，AI辅助的访谈后编辑能够在不损害数据质量的情况下有效降低隐私风险，但仍需指出若干局限。首先，样本的多样性虽经功效分析确定，且足以检测我们关注的主要效应，但总体而言仍较为有限（N = 188）。我们仅选取了来自美国的参与者。更为多样化的参与者群体将有助于通过跨文化效应丰富我们的研究结果，并开展更细致的亚组分析，例如考察人口统计信息、隐私规范、收入水平或使用人工智能工具的过往经验方面的差异。其次，我们的评估仅聚焦于一个访谈主题：在求职面试中使用人工智能。尽管这一主题具有现实意义且敏感度适中，但它并不能代表聊天机器人可能应用的所有敏感领域，例如健康、财务或创伤披露。在这些上下文中，参与者的编辑行为和隐私需求可能存在显著差异。最后，本研究所涉及的敏感程度有限。尽管参与者披露了个人身份信息（例如姓名、学校、所属机构），但与涉及高度污名化或高风险信息（例如医疗状况或非法活动）的披露相比，其风险较低。因此，与高风险领域相比，我们可能低估了隐私风险以及隐私控制的价值。

8 研究设计中的伦理考量

无论筛选结果如何，都给予公平报酬。 我们确保无论参与者是否符合主要研究的资格，都能获得公平报酬。通过筛选的参与者

并完成全部任务的参与者获得了 4 美元，与 Prolific 建议的时薪标准一致。未通过筛选的参与者会被自动筛除，但仍会按照平台指南获得补偿（0.14 美元）。这种设计避免了经济胁迫，即使参与者不符合资格，也尊重了他们的时间。

关于大语言模型的使用和自愿参与的透明度。参与者被明确告知，该聊天机器人由大语言模型驱动。该系统依靠 OpenAI 模型（GPT-4.1 和 GPT-4o-mini）进行对话生成和隐私检测。参与完全出于自愿：用户可在任何阶段选择退出，包括在编辑后拒绝分享原稿转录文本，同时仍可获得全额报酬，如分阶段同意流程所示。我们还披露了 OpenAI 服务的使用情况，并告知参与者有关合同规定的数据处理政策（例如，用户数据不会用于模型训练，数据存储时间最长为 30 天），以增强透明度和信任。

安全的数据存储和仅限研究用途。 所有收集的数据均存储在加密的 AWS S3 存储空间（SSE-S3）中，并实施严格的访问控制；只有获授权的研究人员才能访问参与者数据。会话状态会定期序列化并安全上传至 S3，以防止数据丢失，同时确保保密性。所有数据仅用于研究目的，并以汇总形式进行分析。

9 结论

我们开展了首项针对基于聊天机器人的访谈隐私控制的研究，重点考察了两种访谈后编辑功能：（1）自由编辑；（2）AI辅助编辑。我们的AI辅助编辑流程显著减少了残留的个人身份信息，同时没有降低回答质量或参与度。主观感受并未完全反映实际行为：不同条件下，受试者的舒适度和讨论敏感话题的意愿相近；然而，AI辅助建议将模糊的担忧转化为具体且影响显著的编辑，同时保留了信息的丰富性。

设计启示十分明确：通过行内的跨度级高亮，使风险变得可感知且可操作；优先采用可逆替换和抽象操作；保持建议的保守性与稳定性；并采用领域调优分类体系，优先处理高风险实体。这些选择有助于支持自主权，使用户能够继续掌控整个过程，同时由系统引导其将精力投入到真正有助于保护隐私的编辑上。

我们还观察到收益存在个体差异，这表明可以采用自适应策略（例如，当早期回答中充斥个人身份信息时，使用更强的默认设置或更早地显示提示词）。未来的研究应检验更多领域和人群，提高对高风险跨度的召回率，并探索能够保留用户能动性的实时变体。我们的研究结果，加上日益增多的证据——表明面向用户且可感知的控制能够最大限度地减少不必要的披露——推动了在研究型和产品型聊天机器人中建立实用的隐私保护设计层。

综上所述，我们的研究将事后由用户手动编辑并辅以 AI 辅助提示，定位为传统访谈流程中隐私控制方法的实用补充，并为人机交互研究人员提供了一条可行路径，使其能够将隐私保护设计融入聊天机器人的工作流程，以更少侵入的方式实现这一目标，同时不牺牲丰富且可供分析的数据。

参考文献

[1] Alessandro Acquisti、Idris Adjerid、Rebecca Balebako、Laura Brandimarte、Lor- rie Faith Cranor、Saranga Komanduri、Pedro Giovanni Leon、Norman Sadeh、Florian Schaub、Manya Sleeper、Yang Wang 和 Shomir Wilson。2017年。《隐私与安全助推：理解并协助用户在线选择》。ACM计算调查。 50, 3, 44 2017 41 10 1145 3054926 文章（8月），[2] 2015 页码。doi: . / Alessandro Acquisti、Laura Brandimarte 和 George Loewenstein。。信息时代的隐私与人类行为。《科学》347，6221（2015），509–514。[3] Nazanin Andalibi、Margaret E. Morris 和 Andrea Forte。2018年。试探水温、发送线索：社交媒体上对受社会污名化经历的间接披露。ACM人机交互论文集2，计算机支持的协同工作，文章19（2018年11月），23页。doi:10.1145/3274288[4] Theo Araujo、Jef Ausloos、Wouter van Atteveldt、Felicia Loecherbach、Judith Moeller、Jakob Ohme、Damian Trilling、Bob van de Velde、Claes De Vreese 和 Kasper Welbers。2022年。OSD2F：开源数据捐赠框架。计算传播研究4，2（2022），372–387。[5] Susanne Barth 和 Menno DT De Jong。2017 年。隐私悖论 —— 考察表达的隐私顾虑与实际在线行为之间的差异——系统性文献综述。《远程信息处理与信息学》34，7（2017），1038–1058。[6] Elettra Bietti。2019年。同意作为通行证：平台权力与信息化转向的局限。《佩斯法律评论》40（2019），310。[7] José Patrício Bispo Júnior。2022年。定性健康研究中的社会赞许偏差。《公共卫生杂志》56（2022），101。[8] Andrea Caputo。2017年。自我报告的幸福感测量中的社会赞许偏差：一项在线调查的证据。《心理学大学》16（2017年8月）。doi:10.11144/Javeriana.upsy16-2.sdsw [9] Nicholas Carlini、Florian Tramer、Eric Wallace、Matthew Jagielski、Ariel Herbert- Voss、Katherine Lee、Adam Roberts、Tom Brown、Dawn Song、Ulfar Erlingsson、Alina Oprea 和 Colin Raffel。2021年。从大语言模型中提取训练数据。arXiv: 2012 07805 [ ] 2012 07805 。cs.CR [10] 2001 https://arxiv.org/abs/ 。Justine Cassell。。具身对话代理：用户界面中的表征与智能。《AI杂志》22，4（2001年12月），67。doi:10.1609/ aimag.v 22 4 1593 i 。[11] Hanbyul Choi、Jonghwa Park 和 Yoonhyuk Jung。2018年。隐私疲劳在在线隐私行为中的作用。《计算机与人类行为》81（2018），42 51 10 1016 2017 12 001 –。doi: . /j.chb. . . Sunny Consolvo、Ian E. Smith、[12] Tara Matthews、Anthony LaMarca、Jason Tabert 和 Pauline Powledge。2005年。向社会关系披露位置：人们希望分享什么、何时分享以及为何分享。在SIGCHI计算系统人为因素会议论文集（美国俄勒冈州波特兰）（CHI '05）中。计算机协会，纽约州纽约市， 81 90 10 1145 1054972 1054985 –。doi: . / 。[13] 2021 Rachel Cummings、Gabriel Kaptchuk 和 Elissa M. Redmiles。。“我需要更好的描述”：对用户差分隐私预期的调查。在2021年ACM SIGSAC计算机与通信安全会议论文集（线上活动，大韩民国）（CCS '21）中。计算机协会，美国纽约州纽约市，3037–3052。doi:10.1145/3460120. 3485252[14] Yang Deng、Wenxuan Zhang、Wai Lam、See-Kiong Ng 和 Tat-Seng Chua。2023年。面向大语言模型驱动对话代理的即插即用策略规划器。arXiv预印本 arXiv:2311.00262（2023）。[15] Kathleen Kara Fitzpatrick、Alison Darcy 和 Molly Vierhile。2017年。使用全自动对话代理（Woebot）为具有抑郁症和焦虑症症状的年轻成年人提供认知行为疗法：一项随机对照试验。《JMIR心理健康》4，2（2017年6月6日），e19。doi:10.2196/mental. 7785[16] Blandine French、Camilla Babbage、Katherine Bird、Lauren Marsh、Mirabel Pelton、Shireen Patel、Sarah Cassidy 和 Stefan Rennick-Egglestone。2024年。基于网络的研究中的数据完整性问题：一个普遍挑战的机构案例。《JMIR心理健康》 11 16 2024 58432 10 2196 58432（9月），e 。[17] 2023 doi: . / Yao Fu、Hao Peng、Tushar Khot 和 Mirella Lapata。。通过自我博弈和从人工智能反馈中进行上下文学习改进语言模型谈判。arXiv: 2305 10142 [ ] 2305 10142 。cs.CL https://arxiv.org/abs/ 。[18] Jonathan Gratch、Gale M. Lucas、Aisha Aisha King 和 Louis-Philippe Morency。2014年。这只是一台计算机：人机代理交互对临床访谈的影响。在2014年自主代理与多代理系统国际会议论文集（法国巴黎）（AAMAS '14）中。自主代理与多代理系统国际基金会，南卡罗来纳州里奇兰，85–92。[19] Herbert Paul Grice。1975年。逻辑与会话。《句法与语义》3（1975），43–58。[20] Thomas Groß。2020年。互联网用户信息隐私顾虑量表（IUIPC）的效度与信度 [扩展版]。arXiv:2011.11749 [cs.HC] https://arxiv.org/abs/2011.11749

[21] Ece Gumusel。2024年。对话式聊天机器人中的用户隐私顾虑文献综述：一种社会信息学方法：《信息科学与技术年度评论》（ARIST）论文。《美国信息科学与技术学会会刊》 76 05 2024 121 154 10 1002 24898 （），–。doi: . /asi。[22] K. Handa、M. Stern、S. Huang、J. Hong、E. Durmus、M. McCain、G. Yun、A. Alt、T. Millar、A. Tamkin、J. Leibrock、S. Ritchie 和 D. Ganguli。2025年。介绍Anthropic Interviewer：1,250名专业人士告诉我们的与人工智能协作经验。https://www.anthropic.com/research/anthropic-interviewer。存档于 https://web.archive.org/web/20251204184855/https://www.anthropic.com/research/anthropic-interviewer。[23] David A Herz、Janice E Looman 和 Shirley Kane Lewis。1992年。知情同意：这是一个神话吗？《神经外科》30，3（1992），453–458。doi:10.1227/00006123-199203000- 00028[24] Aashita Jain。2025年。2024年至2025年的人工智能与隐私：拥抱全球法律发展的未来。https://cloudsecurityalliance.org/blog/2025/04/22/ai-and-privacy-2024-to-2025-embracing-the-future-of-global-legal-developments。https://cloudsecurityalliance.org/blog/2025/04/22/ai-and-privacy-2024-to-2025-embracing-the-future-of-global-legal-developments 访问日期：2025年9月。[25] Zoe Zhiqiu Jiang。2024年。向人工智能自我披露：人机交互中的信任与脆弱性悖论。arXiv:2412.20564 [cs.HC] https: //arxiv.org/abs/ 2412 20564 。[26] Ruogu Kang、Laura Dabbish、Nathaniel Fruchter 和 Sara Kiesler。2015年。“我的数据到处流转”：用户对互联网的心智模型及其对隐私与安全的影响。在第十一届可用隐私与安全研讨会（SOUPS 2015）中。USENIX 协会，渥太华，39–52。https://www.usenix.org/ conference/soups2015/proceedings/presentation/kang [27] Jane Kaye、Edgar A Whitley、David Lund、Michael Morrison、Harriet Teare 和 Karen Melham。2015年。动态知情同意：面向21世纪研究网络的患者界面。《欧洲人类遗传学杂志》23，2（2015），141–146。[28] Sunnie S. Y. Kim。2024年。通过透明度与可解释性建立对人工智能的适当信任。在CHI计算机系统人因大会扩展摘要（美国夏威夷州檀香山）（CHI EA ’24）中。计算机协会，美国纽约州纽约市，文章433，6页。doi:10.1145/3613905.3638184[29] Vivian Lai、Samuel Carton、Rajat Bhatnagar、Q. Vera Liao、Yunfeng Zhang 和 Chenhao Tan。2022年。通过条件式委派实现人机协作：内容审核案例研究。在2022年CHI计算机系统人因大会论文集（美国路易斯安那州新奥尔良）（CHI ’22）中。计算机协会，美国纽约州纽约市，文章54，18页。doi:10.1145/3491102.3501999[30] Carl A. Latkin、Catie Edwards、Melissa A. Davey-Rothwell 和 Karin E. Tobin。2017年。巴尔的摩马里兰州城市物质使用者中，社会赞许偏差与健康、物质使用及社会网络因素自我报告之间的关系。《成瘾行为》73（2017），133–136。doi:10.1016/j.addbeh. 2017.05.005[31] Yi-Chieh Lee、Naomi Yamashita、Yun Huang 和 Wai Fu。2020年。“我听见你，我感受你”：通过聊天机器人鼓励深度自我披露。在2020年CHI计算机系统人因大会论文集（美国夏威夷州檀香山）（CHI ’20）中。计算机协会，美国纽约州纽约市，1–12。doi:10.1145/3313831.3376175[32] Guohao Li、Hasan Abed Al Kader Hammoud、Hani Itani、Dmitrii Khizbullin 和 Bernard Ghanem。2023年。CAMEL：用于探索大语言模型社会“心智”的沟通型智能体。在第37届神经信息处理系统大会论文集（美国路易斯安那州新奥尔良）（NIPS ’23）中。Curran Associates Inc.，纽约州雷德胡克，文章2264，18页。[33] Tianshi Li。2026年。智能体型大语言模型作为强大的去匿名化工具：Anthropic Interviewer 数据集参与者的重新识别。arXiv:2601.05918 [cs.CR] https://arxiv.org/abs/ 2601 05918 。Laura H Lind、Michael F Schober、Frederick G Conrad 和 Heidi Reichert。2013年。为什么调查受访者在计算机提问时会披露更多信息？《公众舆论季刊》77，4（2013），888–935。[34] Steven E Locke、Hollis B Kowaloff、Robert G Hoff、Charles Safran、Mark A Popovsky、Deborah J Cotton、Dianne M Finkelstein、Peter L Page 和 Warner V Slack。1992年。用于筛查献血者艾滋病病毒传播风险的计算机辅助访谈。《美国医学会杂志》268，10（1992），1301–1305。[36] Gale M. Lucas、Jonathan Gratch、Aisha King 和 Louis-Philippe Morency。2014年。这只是一台计算机：虚拟人提高披露意愿。《计算机与人类行为》 37 2014 94 100 10 1016 2014 04 043 （），–。doi: . /j.chb. . . [37] Ewa Luger 和 Abigail Sellen。2016。。“就像拥有一个非常糟糕的个人助理”：用户期望与对话代理体验之间的鸿沟。在2016年CHI计算机系统人因大会论文集（美国加利福尼亚州圣何塞）（CHI ’16）中。计算机协会，美国纽约州纽约市，5286–5297。doi:10.1145/2858036.2858288[38] Ying Ma、Shiquan Zhang、Dongju Yang、Zhanna Sarsenbayeva、Jarrod Knibbe 和 Jorge Goncalves。2025年。使用大语言模型提高人们对社交媒体照片中位置信息漏洞的认识。在2025年CHI计算机系统人因大会论文集

25）计算 科统中的人为因素（CHI纽约市，文章238，14页。doi:10.1145/3706598.3714074[39] Michelle Madejski、Maritza Lupe Johnson 和 Steven Michael Bellovin。2011。在线社交网络隐私设置的失效。（2011）。[40] Alice E Marwick 和 danah boyd。2014。网络化隐私：青少年如何在社交媒体中协商情境。《新媒体与社会》16，7（2014），1051–1067。arXiv:https://doi.org/10.1177/1461444814543995 doi:10.1177/1461444814543995 [41] Brian James McInnis、Ramona Pindus、Daniah Kareem、Savannah Gamboa 和 Camille Nebeker。2024。探索知情同意的未来：应用服务设计方法。ACM人机交互论文集 8，CSCW1，文章 53 2024 31 10 1145 3637330 （4月），页码。doi: 。/ Sara Memarian[42]、Onkar S Malgonde 和 Dan J Kim。2025。揭开隐私悖论：在线隐私顾虑与披露行为差异背后因素的全面综述。《信息系统前沿》（2025），1–22。[43] 微软。2025。Presidio：数据保护与去标识化软件开发工具包。https: //microsoft.github.io/presidio/.[44] Niloofar Mireshghallah、Maria Antoniak、Yash More、Yejin Choi 和 Golnoosh Farnadi。2024。Trust No Bot：发掘真实情境下人与大语言模型对话中的个人信息披露。arXiv:2407.11438 [cs.CL] https://arxiv.org/abs/2407. 11438[45] Justin M Mittelstädt、Julia Maier、Panja Goerke、Frank Zinn 和 Michael Hermes。2024。大语言模型在社会情境判断方面可以胜过人类。《科学报告》14，1（2024），27449。[46] Emma Nicol、Jo Briggs、Wendy Moncur、Amal Htait、Daniel Paul Carey、Leif Azzopardi 和 Burkhard Schafer。2022。揭示在线个人信息中的累积风险：一项数据叙事研究。ACM人机交互论文集 6, 2, 323 2022 25 10 1145 3555214 CSCW文章（11月）[47]，页码。doi: 。/ OpenAI。。GPT- o mini：推动高性价比智能发展。https://openai. com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. GPT-4o mini的公告和定价详情。[48] OpenAI。2025。在 API 中推出 GPT-4.1。https://openai.com/index/gpt-4-1/。GPT-4.1、GPT-4.1 mini 和 GPT-4.1 nano 的发布公告。[49] OpenAI。2025。OpenAI。https://www.openai.com。通过 OpenAI 网站及其出版物访问。[50] Aswati Panicker、Novia Nurain、Zaidat Ibrahim、Chun-Han (Ariel) Wang、Se- ung Wan Ha、Elizabeth Kaziunas、Maria K Wolters 和 Chia-Fang Chung。2024。以人为中心的设计中的欺诈行为：定性人机交互研究的集体策略与未来议程。载于《CHI计算机系统人因大会扩展摘要集》（美国夏威夷州檀香山）（CHI EA ’24 469，5 10 1145 3613905）。计算机协会，美国纽约州纽约市，文章页码。doi: 。/ 。[51] Aswati Panicker、Novia Nurain、Zaidat Ibrahim、Chun-Han (Ariel) Wang、Se- ung Wan Ha、Yuxing Wu、Kay Connelly、Katie A. Siek 和 Chia-Fang Chung。2024。理解在线定性研究中的欺诈行为：研究者视角。载于《2024年CHI计算机系统人因大会论文集》（美国夏威夷州檀香山）（CHI ’24）。计算机协会，美国纽约州纽约市，文章 824, 17 10 1145 3613904 3642732 页码。doi: 。/ 。Delroy L. Paulhus[52]。。章节——反应偏差的测量与控制。载于《人格与社会心理态度测量》，John P. Robin- son、Phillip R. Shaver 和 Lawrence S. Wrightsman（编）。学术出版社，17–59。doi:10.1016/B978-0-12-590241-0.50006-X[53] Matthew D. Pickard、Catherine A. Roster 和 Yixing Chen。2016。揭示个人访谈中的敏感信息：在什么条件下，与人还是虚拟化身交谈更容易自我披露？《计算机与人类行为》65（2016），23 30 10 1016 2016 08 004 – 。doi: 。/j.chb. 。 。Wendy L Richman、Sara Kiesler、Suzanne Weisband 和 Fritz Drasgow。1999。计算机施测问卷、传统问卷和访谈中的社会赞许性偏差：一项荟萃分析研究。《应用心理学杂志》84，5（1999），754。[55] Piero Romare、Farzaneh Karegar 和 Simone Fischer-Hübner。2025。迈向物联网 TAP 的可用隐私管理：推导隐私群组与偏好档案。arXiv: 2511 11209 [ ] 2511 11209 。cs.HC [56] https://arxiv.org/abs/ 。Jayshree Sarathy、Sophia Song、Audrey Haque、Tania Schlatter 和 Salil Vadhan。2023。别看数据！差分隐私如何重塑数据科学实践。载于《2023年CHI计算机系统人因大会论文集》（德国汉堡）（CHI ’23）。计算机协会，美国纽约州纽约市，文章164，19页。doi:10.1145/3544548. 3580791[57] Advait Sarkar。2025。这可能是人工智能写的：知识工作中一种阶级歧视性侮辱语的诞生。载于《CHI计算机系统人因大会扩展摘要集》（CHI EA ’25）。计算机协会，美国纽约州纽约市，文章621，12页。doi:10.1145/3706599. 3716239[58] Florian Schaub、Rebecca Balebako、Adam L. Durity 和 Lorrie Faith Cranor。2015。有效隐私告知的设计空间。载于第十一届 USENIX 可用隐私与安全会议论文集（加拿大渥太华）（SOUPS ’15）。USENIX 协会，美国，1–17。

[59] Ruosi Shao。2023。用于心理健康干预的共情人工智能：概念化并考察人工共情。载于第二届以共情为中心的设计研讨会论文集（德国汉堡）（EmpathiCH ’23）。计算机协会，美国纽约州纽约市，文章4，6页。doi:10.1145/3588967.3588971[60] Katie Shilton。2012。参与式个人数据：信息科学面临的一项新兴研究挑战。《美国信息科学与技术学会杂志》63，10（2012），1905–1915。[61] Steven Siddals、John Torous 和 Astrid Coxon。2024。“恰好是最合适的选择”：生成式人工智能聊天机器人用于心理健康的体验。npj心理健康研究 3，1（2024），48。[62] Daniel J Solove。2013。引言：隐私自我管理与同意困境。《哈佛法律评论》126，7（2013），1880–1903。[63] BM Stanley、DJ Walters 和 GJ Maddern。1998。知情同意：提供多少信息才足够？《澳大利亚和新西兰外科学杂志》68，11（1998），788–791。[64] Arkadiusz Stopczynski、Riccardo Pietri、Alex Pentland、David Lazer 和 Sune Lehmann。2014。传感器驱动的人类数据收集中的隐私：从业者指南。arXiv预印本 arXiv:1403.5299（2014）。[65] Frederic D Stutzman、Ralph Gross 和 Alessandro Acquisti。2013。沉默的倾听者：Facebook 上隐私与披露的演变。《隐私与保密杂志》4，2（2013），2。[66] Latanya Sweeney。2002。k-匿名性：一种隐私保护模型。《国际不确定性、模糊性与知识型系统杂志》10，05（2002），557–570。[67] Roger Tourangeau 和 Ting Yan。2007。调查问卷中的敏感问题。《心理学通报》133，5（2007），859。[68] Blase Ur、Pedro Giovanni Leon、Lorrie Faith Cranor、Richard Shay 和 Yang Wang。2012。智能、有用、可怕、令人毛骨悚然：人们对在线行为广告的看法。载于第八届可用隐私与安全研讨会论文集。1–15。[69] Ge Wang、Jun Zhao、Max Van Kleek 和 Nigel Shadbolt。2023。赋能的12种方式：为儿童数字自主性进行设计。载于《2023年CHI计算机系统人因大会论文集》。1–27。[70] Liuping Wang、Dakuo Wang、Feng Tian、Zhenhui Peng、Xiangmin Fan、Zhan Zhang、Mo Yu、Xiaojuan Ma 和 Hongan Wang。2021。Cass：为在线健康社区构建社会支持聊天机器人。《ACM人机交互汇刊》5，CSCW1（2021），1–31。[71] Yang Wang、Gregory Norcie、Saranga Komanduri、Alessandro Acquisti、Pedro Giovanni Leon 和 Lorrie Faith Cranor。2011。“我按下分享按钮的那一刻就后悔了”：一项关于 Face book 上后悔情绪的定性研究。载于第七届可用隐私与安全研讨会论文集。1–16。[72] Suzanne Weisband 和 Sara Kiesler。1996。计算机表单中的自我披露：荟萃分析及其启示。载于 SIGCHI计算系统人为因素会议论文集。3–10。[73] Ziang Xiao、Tiffany Wenting Li、Karrie Karahalios 和 Hari Sundaram。2023。告知不知情者：利用人工智能驱动的聊天机器人改进在线知情同意阅读。载于《2023年CHI计算机系统人因大会论文集》。1–17。[74] 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含开放式问题的对话式调查。《ACM计算机与人类交互汇刊》（TOCHI）27，3（2020），1–37。[75] Xiao Zhan、Juan Carlos Carrillo、William Seymour 和 Jose Such。2025。恶意的基于大语言模型的对话式人工智能会诱使用户披露个人信息。arXiv预印本 arXiv:2506.11680（2025）。[76] Zhiping Zhang、Michelle Jia、Hao-Ping Lee、Bingsheng Yao、Sauvik Das、Ada Lerner、Dakuo Wang 和 Tianshi Li。2024。“这很公平”，还是如此？探究用户使用基于大语言模型的对话代理时如何权衡披露风险与收益。载于《2024年CHI计算机系统人因大会论文集》。1–26。[77] Zhiping Zhang、Chenxinran Shen、Bingsheng Yao、Dakuo Wang 和 Tianshi Li。2025。大语言模型（LLM）的秘密使用。《ACM人机交互汇刊》9，2（2025），1–26。[78] Jijie Zhou、Eryue Xu、Yaoyao Wu 和 Tianshi Li。2025。Rescriber：由较小大语言模型驱动、由用户主导的数据最小化，适用于基于大语言模型的聊天机器人。载于《2025年CHI计算机系统人因大会论文集》。1–28。

A 完整访谈流程

A.1 筛选（资格筛选）问题

以下我们通过题目符合条件的选项加粗来标示纳入标准。参与者可以看到这些选项，但不会被告知必须选择这些选项才能符合研究资格。

(1) 您是否年满18岁？

- **是**
- 否 (2) 您能流利地说英语并理解英语吗？
- **是**
- 否 (3) 以下哪项最能描述您在求职面试中使用人工智能的情况？（请选择一项）
- **我曾在求职面试期间使用人工智能（例如，在现场或远程访谈中）。**
- **我曾使用人工智能准备求职面试（例如，进行模拟面试、问答练习、起草回答）。**
- **我曾使用人工智能准备求职面试，并在求职面试期间使用人工智能。**
- 我未曾在求职面试中使用人工智能。
- 其他

A.2 任务后调查：通用条目（所有组别）

（1）在访谈期间或之后，你是否对自己的隐私有任何顾虑？如果有，请详细说明。• 开放式回答（自由文本）。（2）在会话过程中，是否有任何时刻你因隐私顾虑而犹豫是否回答？如果有，请详细说明。

• 开放式回答（自由文本）。（3）与聊天机器人分享个人信息时，我感到舒适。

- 5点李克特同意量表（见李克特量表说明小节）。

（4）我希望在下一次与聊天机器人的访谈中讨论敏感话题。

- 5点李克特同意量表（见李克特量表说明小节）。

A.3 任务后调查：人口统计条目（所有组）

（1）您的年龄组是？

- 18–29
- 30–39
- 40–49
- 50–59
- ≥ 60
- 不愿透露

(2) 您的性别是什么？

- 男性
- 女性
- 非二元性别
- 不愿透露

(3) 您的种族和族裔是什么？ ?

- 黑人
- 白人
- 亚洲人
- 西班牙裔
- 美洲原住民
- 不愿透露

(4) 您的教育程度是什么?
- 无文凭或低于十二年级
- 高中
- 大学学历
- 副学士学位
- 学士学位
- 研究生或专业学位

A.4 任务后调查：自由编辑组专属条目

（仅显示给自由编辑组 .)

（1）您在会话中删除或修改了哪些内容？为什么？• 开放式回答（自由文本）。（2）对于分享最终编辑版本，您的舒适度如何？请说明您的舒适度以及原因。• 开放式回答（自由文本）。（3）编辑功能提升了我的隐私控制感。

- 5点李克特量表同意度量表（见李克特量表描述小节）。

（4）我能够清楚地理解如何使用编辑功能来保护我的隐私。

- 5点李克特量表同意度量表（见李克特量表描述小节）。

（5）自由编辑让我对自己分享的内容更有信心。

- 5点评分的李克特同意度量表（见“李克特量表”说明小节）。

A.5 任务后调查：AI辅助编辑组专属条目

（仅向 AI编辑组显示 .)

（1）你根据 AI建议在会话中删除或更改了哪些内容？为什么？• 开放式回答（自由文本）。（2）你对分享自己的最终编辑版本感到多舒适？请说明你的舒适度以及原因。• 开放式回答（自由文本）。（3）AI辅助编辑提升了我对隐私控制的感受。

- 5点评分的李克特同意度量表（见“李克特量表”说明小节）。（4）我能够清楚地理解 AI建议如何帮助我保护隐私。
- 5点评分的李克特同意度量表（见“李克特量表”说明小节）。（5）AI辅助编辑让我对自己分享的内容更有信心。

- 5点李克特同意度量表（参见李克特量表说明小节）。

A.6 舒适度与感知隐私控制的五点李克特量表

所有关于舒适度的李克特式条目（例如，“我在与聊天机器人分享个人信息时感到舒适。”）以及感知隐私控制（例如，“编辑功能提升了我的隐私控制感。”，“AI辅助编辑提升了我的隐私控制感。”）均使用相同的5点同意量表:

- 1 =非常不同意
- 2 =不同意
- 3 =中立
- 4 =同意
- 5 = 非常同意

A.7 招募帖：第 1 部分（预筛选）：预筛选问卷：求职面试中的AI使用（根据双盲政策隐去个人信息）

关于求职面试中的人工智能研究的 1 道题预筛选问卷（~1 分钟），由 [大 学] 开展。对**主要研究（~15 分钟）**感兴趣吗？请完成此预筛选问卷，以便获得考虑资格。

关于主要研究（适用于感兴趣者）。如果您对**主要研究**感兴趣，该研究包括一次关于在求职面试中使用（或不使用）人工智能的**纯文本聊天机器人访谈**，以及一份简短的后续问卷。**持续时间：~15分钟**。**主要研究的预计报酬**为**$3**，通过 Prolific**发出邀请并不保证；**资格取决于预筛选问卷的回答和目标配额。完成此预筛选问卷**表示您有兴趣**被纳入考虑。**主办方与团队**。 [大 学] 研究人员：[研究人员]**目的**。确定您是否有资格参加一项关于人们如何使用或选择不使用人工智能进行求职面试的研究。**您需要做什么**。回答一**道多项选择题**，内容关于您在求职面试中使用人工智能的情况。**时间与付款**。 ~1分钟；按照 Prolific 上列出的费率支付。**自愿参与**。参与是自愿的；您可以随时退出。**数据与隐私**。我们仅记录您的**Prolific ID** 和您的一次回答。您在本研究中的参与情况将予以保密；不会发布或展示个人标识符。**联系方式**。有关研究的问题： [研究人员2

]

[电子邮件地址2]； [研究人员1] [电子邮件地址1].**您的权利**。 有关您作为参与者的权利的问题： [大学] 人类研究部，电话： [电话号码], 电子邮件： [电子邮件地址3] （您可以匿名致电）。**经费**。 [大学] . 版本日期： [日期] .

A.8 招募帖：关于在求职面试中使用人工智能的聊天机器人访谈（根据双盲政策隐去个人信息）

赞助方与团队。 [大学]研究人员：[研究人员]

标题。 关于在求职面试中使用人工智能的聊天机器人访谈

目的。 了解人们在分享个人故事或解决问题时如何与人工智能聊天机器人互动。

您将要做什么。（1）提供同意。（2）完成一项关于在求职面试中使用人工智能的**纯文本聊天机器人访谈**（您**可以跳过任何问题** 或随时停止）。（3）回答一份后测问卷，包括收集您对系统的反馈以及年龄和性别等人口统计信息。（4）您将被重定向至感谢页面，以完成研究。

持续时间与报酬。 ~20分钟。通过 Prolific 支付 4 美元；具体金额将在 Prolific 研究页面上显示。提交内容将及时审核。

自愿参与。参与完全自愿；您可以在提交前随时退出研究。

数据与隐私。您在本研究中的参与信息将予以保密。只有研究团队会查看您的信息；个人标识符不会被发表或公开展示。我们会收集您的 Prolific ID，以关联您的参与记录并处理报酬。

联系方式。有关研究的问题：[2] [1] [1] [2] 研究人员电子邮件地址；研究人员电子邮件地址。

您的权利。有关您作为参与者的权利的问题：[大学] 人类研究部，电话：[PHONENUMBER], 电子邮件地址：[EMAIL3]（您可以匿名致电）。

经费。 [大学]。版本日期：[DATE]。

A.9 第1组：教育背景

主要提示词。可以告诉我你的教育背景吗——你在大学里学了什么?

条件性后续问题。

Q1_F1 你什么时候开始攻读学位，什么时候完成学业? Q1_F2 你的大学位于哪里？ Q1_F3 在你攻读学位期间，谁是你觉得教给你最多的人？说出一个人的名字，有时能帮助你更具体地回顾自己的经历。

A.10 第2组：当前工作与访谈路径

Mainprompt. 我很想了解你目前的工作，以及你是如何通过求职面试进入这一行的?

条件性后续问题。Q2_F1 你什么时候开始做当前工作，又是什么时候进行访谈的？ Q2_F2 你目前的职位名称是什么？ Q2_F3 你目前的雇主或公司是什么?

A.11 第 3 组：求职面试前后使用人工智能的具体情况

主要提示词。 你能具体讲讲，之前在求职面试前和面试过程中，你曾在什么时候使用过人工智能吗?

条件性后续问题。Q3_F1 你具体在求职面试流程的哪个阶段使用了人工智能？ Q3_F2 你使用了哪些人工智能工具，分别用于哪些任务？ Q3_F3 你如何使用人工智能？这带来了什么不同?

A.12 第 4 组：险些出问题／紧张时刻

主要提示词。 你在求职过程中是否曾有过这样的时刻：使用人工智能让你感到紧张，担心可能会带来问题？可以跟我讲讲吗?

条件性后续问题。

Q4_F1 你还记得这件事发生在什么时间吗？ Q4_F2 你使用了哪些人工智能工具，分别用于哪些任务？ Q4_F3 能否详细说说这件事当时对你产生了什么影响？这是真正的问题，还是更多只是担忧?

A.13 第5组：边界与伦理反思

主要提示词。 你是否曾怀疑过，自己在访谈中使用人工智能是否完全恰当？如果有，是什么让你产生了这种感觉?

条件性后续问题。

Q5_F1 你能否跟我说说，是什么让你在那种情况下使用人工智能？ Q5_F2 当你想到使用人工智能可能会越界时，你在担心什么？ Q5_F3 在求职面试中使用人工智能时，你还进行过哪些伦理方面的反思?

A.14 第6组：隐秘或私下使用人工智能

主提示词。 你是否曾在求职面试中使用人工智能，但不愿公开告诉他人——例如雇主、家人、朋友或同事?

条件性后续问题。Q6_F1 你使用人工智能躲避某人的那次事件发生在什么时候？Q6_F2 你当时在躲避谁？Q6_F3 你试图向他们隐瞒哪些人工智能用途?

Q6_F4 你为什么觉得那是你不愿公开分享的事情?

B 大语言模型提示词

B.1 执行器提示词（编排器）

指示模型扮演“执行器”，引出具体的个人故事、围绕当前问题展开，并输出结构化 *JSON*。

当前问题："<当前问题或不适用>" 剩余问题：["..."; "..."]允许的_操作：[提出_后续问题、请求_澄清、总结_问题]

你是执行器。你的工作是针对当前问题引出具体的个人故事。硬性规则：- 只围绕当前问题展开；不要引入其他预设问题。- 主问题：表达简洁、自然，每次只问一个重点明确的后续问题；语气温暖、好奇、中立。- 目标是了解：时间/地点/人物/任务/行动/结果 + >=2 深度要点（权衡/困难/失败的尝试/反思）。- 当你认为已达到要求时，请先提出一个 2-3 行的总结，再继续下一个问题。

仅输出 JSON：{ "action": "提出_后续问题" | "总结_问题" | "请求_澄清" | "结束", "question_id": "<ID 或文本>", "utterance": "一个自然的问题或简短摘要", "notes": ["可选的已提取事实"]}

B.2 后续问题覆盖审核器 (PSS)

更严格的覆盖范围审计器；“已覆盖”要求该后续问题已被明确提出并得到回答；恰好选择下一个尚未覆盖的后续问题。

你是后续问题覆盖审核器。目标：针对当前问题，检查用户迄今为止的会话是否已经覆盖每一项必需的后续问题。

当前问题："<问题>" 后续问题：

```
[
  {"id":"Qx_Fy","提示词":"...","keywords":[...]},
  ...
]
```

规则：
- “已覆盖”是指明确提出并得到回答的特定后续问题，而不只是顺带提到了相关信息。- 重点判断每个后续提示词是否确实作为直接问题提出，无论其他地方是否提供了相关信息。只有在有证据表明助手明确提出了该特定问题时，后续问题才算“已覆盖”。

后续问题 .

- 将所有最近的对话轮次全部用作证据（而不仅仅是最后一条消息）。- 如果所有后续问题都已覆盖，则 => verdict ="ALLOW_NEXT_QUESTION"。- 否则 => verdict = "REQUIRE_MORE"，并恰好选择一个仍未覆盖的下一项后续问题（优先采用较早的顺序）

仅输出有效 JSON（无 Markdown，不要添加多余逗号）：{ "question_id": "<text>", "coverage_map": [{ "id": "Qx_Fy", "covered": true|false, "evidence": "单个引文或空字符串" }], "next_followup_id": "Qx_Fy" | null, "next_followup_prompt": "字符串 |null", "verdict": "ALLOW_NEXT_QUESTION" | "REQUIRE_MORE", "confidence": 0.0-1.0, "notes": "简短说明" }

B.3 问题存在性/形式审核器

确保助手消息中恰好包含一个切题的后续问题 - 后续提问模式；如果内容偏离主题或格式不正确，可以请求重新生成。你是问题形式审核器。检查助手的最新消息，判断：(1) 是否包含问题及其形式；(2) 内容是否与当前问题的主题一致。

当前问题："<question>" 当前主题关键词：[kw...]其他主题关键词（避免）：[kw...]

仅返回此 JSON 对象，使用以下格式：{ "hasQuestion": true|false, "reason": "简短说明", "confidence": 0.0-1.0, "shouldRegenerate": true|false }

规则：- 在后续提问模式=true 下，消息必须且只能包含一个以“？”结尾的疑问句（不得连续提出多个问题）。- 如果后续提问模式=false：* 建议提出问题，但简短的总结或过渡消息不必包含问题。* 如果消息明显是总结或过渡内容，则 hasQuestion 可以为 false，shouldRegenerate=false。- 主题一致性：* 问题必须围绕当前主题关键词。* 如果问题似乎引入了另一个预先设定的问题，则将 shouldRegenerate=设为 true。- 对后续提问模式下“必须且只能一个问题”的要求严格把关。- reason 应简短实用。

B.4 简洁存在性审核器（实用工具）

简要检查下一条消息是否恰好是一个切题的问题；返回一个简单的 *ok* 标记。

你是一名简洁的审计器。目标：确保下一条助手消息恰好是一个问题，且与当前问题切题。规则：- 恰好一个疑问句，并以“？”结尾 - 必须与当前问题及其主题提示（如果有）保持一致 - 不得包含多部分问题、列表或开场白 当前问题：“<问题>” 主题提示： <逗号分隔的关键词或（无）>其他主题（避免偏离）： <逗号分隔的关键词或（无）>候选消息：“<助手文本>” 仅输出 JSON：{"ok": true|false, "reason": "简要说明"}

B.5 无经验/拒绝分类器

判断用户是否因缺乏相关经历或拒绝回答而无法提供实质性回答；输出必须为严格的 *JSON* 格式。

你是一个严格的分类器，用于判断用户是否无法针对当前问题提供实质性回答。仅返回有效的 JSON：{"label":"NO_ABLE_ANSWER"|"HAS_ANSWER","reason":"...","reason_type":"no_experience"|"refusal", "evidence":["..."]} 以下任一情况均标记为 NO_ABLE_ANSWER：(A) 没有相关个人经历可分享；(B) 明确拒绝或表示目前无法分享详情（涉及隐私/保密性/保密协议/感到不自在/宁愿不说）。对于类似以下的拒绝措辞，务必标记为拒绝：“抱歉，我无法分享”，“我宁愿不说”，“不愿透露”，“我不能讨论这个”，“这是隐私/机密”，“分享这些让我感到不自在”，“受保密协议约束”。请采取保守判断；只有当用户表示自己能够分享一个具体例子时，才返回 HAS_ANSWER。

B.6 后续问题衔接语润色

围绕给定的后续问题生成简短的陈述式前缀或后缀；绝不改动问题本身。

围绕给定的后续问题编写简短的衔接语。严格规则：- 绝不修改、改述或重新排版该问题。你只需提供“前缀”和“后缀”。- 前缀/后缀中不得提问；必须使用陈述句（任何位置都不得出现“？”）。- 不得以任何方式重述或改述该问题。- 语气：温暖、中立、简洁，符合访谈风格；不使用表情符号；避免重复；不使用 Markdown。- 自然衔接紧邻的上下文（假定使用英文界面）。- 保持简短：前缀 <= 120 个字符，后缀 <= 120 个字符。- 如果上下文涉及敏感内容，请温和回应（例如：“谢谢你愿意分享。”）。- 最终输出应符合语法。仅输出 JSON：{ "prefix": "...", "suffix": "..." }

B.7 单问题重生成器

将助手消息改写为恰好一个切题的疑问句，以满足存在性约束。

你是一名改写助手。请严格生成一个疑问句，并满足以下要求：- 内容严格围绕当前问题的主题，- 针对可能缺失的具体细节（时间/人物/结果、数字、障碍、权衡等），- 表达自然简洁（<=220 个字符），- 以“？”结尾，且不包含其他问号，- 不得添加开场白或解释——只返回问题。

当前问题：“<question>” 当前主题关键词（请勿偏离）： [kw...]原稿用户： <userMessage>原稿助手： <originalResponse> 仅返回单个问题（不要加引号或使用 Markdown）。

B.8 审计引导的问题润色器

根据覆盖范围审计中缺失的后续问题，将下一条消息改写为一个有针对性的问题。

范围审计中缺失的后续问题。
将助手的下一条消息改写为恰好一个
有针对性的问题，以解决该
审计缺口。

约束：
- 严格围绕当前问题。- 有帮助时，请与以下主题提示保持一致（不要引入新主题）：<逗号分隔的关键词>

当前问题："<问题>" 当前主题关键词： [关键词.....]原稿用户： <用户消息>原稿助手： <原始回复>

审核（简要）：判定=<.....>；缺失项=<Qx_Fy>；建议="<提示词文本>"

仅返回单个问题。

B.9 隐私检测提示（个人身份信息分类法）

基于分类法的个人身份信息检测器；分配编号占位符并返回严格的 *JSON*。

你是网络安全和数据隐私领域的专家。现在要求你仅使用以下分类体系，从给定文本中检测个人身份信息：

地址 IP_地址 统一资源定位符 社会安全号码 电话_号码 电子邮件地址 驾驶员_执照 护照_号码 纳税人_识别_号码

ID_编号
NAME
用户名
密钥：密码、通行密钥、API 密钥、加密密钥以及任何其他形式的安全密钥。地理位置：地点和位置，例如城市、省份、国家、国际地区或具名基础设施（例如公交车站、桥梁等）。所属机构：组织名称，例如公共和私营公司、学校、大学、公共机构、监狱、医疗机构、非政府组织、教会等。这包括大学名称，例如“卡内基梅隆大学”“麻省理工学院”“斯坦福大学”等。人口统计_属性：个人的人口统计属性，例如母语、血统、文化遗产、族裔、国籍、宗教或政治团体、胎记、年龄、性取向、性别和生理性别。时间：对特定日期、时间或持续时间的描述。健康_信息：有关个人健康状况、医疗状况、治疗记录和健康保险信息的详细信息。财务_信息：财务详情，例如银行账号、信用卡号、投资记录、薪资信息以及其他财务状况或活动。教育_记录：教育背景详情，包括学业记录、访谈记录、学位和资格证书。对于用户发送给聊天机器人的消息，仅使用上述分类体系识别其中所有个人身份信息。请注意，这些信息应与非公共场合中的真实个人相关，但不一定能够唯一识别该个人。结果应采用可能的最小单位。重要提示：为每个检测到的个人身份信息分配一个带编号的占位符（例如 NAME1、NAME2、EMAIL1 等），编号应在整个对话历史中累计。编号应跨所有对话依次递增，而不仅限于单条消息。例如，如果之前的消息中使用了 NAME1，则下一个姓名应为 NAME2。重复实体检测：如果检测到之前提到过的同一实体（“卡内基梅隆大学”），则使用相同的占位符编号。例如，如果“卡内基梅隆大学”之前被标记为 AFFILIATION1，则同一实体再次出现时仍使用 AFFILIATION1。关键分类规则：- “卡内基梅隆大学”“麻省理工学院”“斯坦福大学”等应归类为所属机构，而非姓名 - 学术领域、专业和学科属于教育_记录 - 只有真实的人名（例如“John Smith”“Sarah Johnson”）才应归类为姓名 - 大学和组织应归类为所属机构

- TIME 必须是**具体的**。仅当文本包含至少一个明确的时间锚点时，才将其标记为 TIME，例如：* 日历日期（如：“2023-05-12”、“05/12/2023”、“2023年5月”、“6月5日”）* 钟表时间（如：“下午3点”、“14:30”）* 带单位的带数字的时长（如：“持续3年”、“超过2个月”、“10天后”）* 具有明确时间语境的年份（如：“2022年”、“自2019年以来”、“2020年至2021年之间”）- 对于没有具体锚点的泛指或疑问性短语，**不要**标记为 TIME，例如：“何时”、“具体时间”、“当时”、“时间”、“时间段”、“这些天”、“那时”。请严格使用以下格式：{ "隐私_问题": true/false, "检测到的_PII": [{ "类型": "PII_类别", " 原始 _ 文本 ": " 找到的精确 _ 文本 _", " 占位符 ": " 类别 1", " 说明 ": "简要_原因" }],"包含_占位符的_文本": "将原始_文本_中的_PII_替换_为_占位符后的文本", "受影响的_文本": "以逗号_分隔的_所有_检测到的_文本_列表_" } 如果未发现隐私问题，请回复：{ "隐私_问题": false, "检测到的_PII": [],"包含_占位符的_文本": "原始_文本_不变", "受影响的_文本": null }

当前用户消息："<消息>"<可选上下文>

B.10 隐私抽象提示词

仅将占位符改写为通用术语；输出映射结果数组。

改写文本以抽象化受保护的信息 ，
且不更改其他部分。
重要：仅操作提供的确切占位符；
不要更改任何其他文本。

例如：
输入： <Text>我毕业于卡内基梅隆大学，年薪六位数。今天在办公室里.....</Text><ProtectedInformation>卡内基梅隆大学,今天</ProtectedInformation>输出 JSON：{"results": [{"protected": "卡内基梅隆大学", "abstracted": "一所声望卓著的大学"}, {"protected": "今天", "abstracted": "最近"}]}

当前输入：
<文本><文本_包含_占位符></文本>要替换的占位符（用逗号分隔，必须完全匹配）：<受保护信息>P1、P2.....</受保护信息><受影响文本><如有，请用逗号分隔列出原文></受影响文本>

请严格使用以下格式：
{"results": [{"protected": "卡内基梅隆大学", "
abstracted": "一所著名大学"},
{"protected": "今天", "abstracted": "最近"}]}

B.11 RQI评估提示词：相关性

B.11.1 相关性定义。

相关性（0-2 分制）：- 0：不相关——与所提问题完全无关（例如，“Yhhchxbxb”这样的胡言乱语）

- 1：有一定相关性 - 没有直接回答问题，但仍提供了有用的信息（例如，问题：“我很想进一步了解这件事。你能想到谁是你在攻读学位期间学到最多的人吗？”回答：“英语专业”；问题：“谢谢你分享这些。你能详细说说是什么让你在那种情况下使用人工智能吗？”回答：“这项任务就是直接分配给我的。”；问题：“谢谢你分享这些。你能详细说说是什么让你在那种情况下使用人工智能吗？”回答：“当时很容易就能用到”）

- 2：相关 - 直接且清楚地回答了所提的问题（例如，问题：“感谢你分享这些信息。你是否曾在求职面试中以不愿公开告诉他人的方式使用过人工智能？”回答：“没有”；问题：“我很想进一步了解这件事。你目前的雇主或公司是什么？”回答：“Matrix Integration”；

问题你的这份工作在哪里？”回答：“新泽西州”）

B.11.2 提示词。

你正在评估答案与问题的相关性。

{相关性_定义}

问题：{question}

答案：{answer}

请评估此答案的相关性，并提供
你的分数，作为具有以下内容的 JSON 对象
格式：
{{
"相关性": <0 ，1，或 2>
}}

仅回复 JSON 对象，不要附加任何文本 .

B.12 RQI 评估提示：具体性

B.12.1 具体性定义。

具体性（0-2分制）：
- 0：仅有笼统描述——仅提供浅层或
抽象描述（e .g．我没有；
我没有任何顾虑）

- 1：具体概念——传达后续信息（例如：我正在申请一个新的校长职位，必须录制一段视频来回答某些问题。我使用人工智能，输入提示词和我的笔记，它帮助我进行整理）

- 2：带有详细示例的具体概念——提供更为详细的描述及具体示例（例如：Matrix Integration（公司名称）；助理校长（职位名称）；新泽西州（地点））

B.12.2 提示词。

你正在评估某个问题答案的具体性。

{具体性_定义}

问题：{question}

答案：{answer}

请评估此答案的具体性，并按照以下格式以 JSON 对象提供评分：

{{ "具体性": <0、1 或 2>}}

仅返回 JSON 对象，不要添加任何其他文本。

B.13 RQI评估提示词：清晰度

B.13.1 清晰度定义。

清晰度（0-2 分制）：- 0：无法辨读的文本 - 胡言乱语或毫无意义的回答（例如，类似“Yhhchxbxb”的胡言乱语）

- 1：不完整句子 - 不完整句子或妨碍理解的语法错误（例如“前一天”）

- 2：表达清晰的回答 - 使用完整句子表达，且没有严重的语法问题（例如：我从本月初开始做现在这份代课教师的工作。我在今年六月参加了访谈。）

B.13.2 提示词。

你正在评估对某个问题的回答是否清晰。

{清晰度_定义}

问题：{question}

回答：{answer}

请评估此回答的清晰度，并按照以下格式以 JSON 对象提供评分：

```
{{
    "Clarity": <0, 1, or 2>
}}
```

请仅回复 JSON 对象，不要附加任何其他文本。

C 参与者人口统计特征

为了将我们的研究发现置于具体背景中，我们报告了参与者的人口统计分布（N=188）。表 5 下方汇总了年龄、性别、种族/族裔和教育程度的人数与百分比。

表 5：参与者人口统计特征（N=188）

领域	类别	n (%)
Age	30-39	61 (32.45%)
	40-49	54 (28.72%)
	18-29	38 (20.21%)
	50-59	27 (14.36%)
	>=60	8 (4.25%)
性别	女性	97 (51.60%)
	Male	88 (46.81%)
	非二元性别	3 (1.59%)
种族/族裔	白人	115 (61.17%)
	黑人	41 (21.81%)
	亚洲人	13 (6.91%)
	西班牙裔	12（6.38%）
	不愿透露	5（2.66%）
	美洲原住民	2（1.07%）
教育程度	学士学位	76 (40.43%)
	研究生或专业学位	45 (23.94%)
	大学学历	28（14.89%）
	高中	20（10.64%）
	副学士学位	18（9.57%）
	无文凭或低于十二年级	1 (0.53%)

D 开源与闭源模型对比分析

在我们的研究中，大语言模型是支撑聊天机器人访谈系统、个人身份信息检测和数据质量 RQI 评估自动化的底层模型。尽管闭源大语言模型的性能更佳，但将数据共享给服务提供商也会引发潜在隐私泄露的担忧。因此，我们开展了实验并参考了以往的研究，阐述在各个部分使用闭源模型的必要性。

D.1 聊天机器人访谈系统

作为研究的主体，为了收集参与者的回答，实施能够在参与者数量庞大时提供流畅交互的模型至关重要。然而，尽管回答质量相近，闭源模型服务与在线运行开源模型之间的响应时间却存在显著差距。

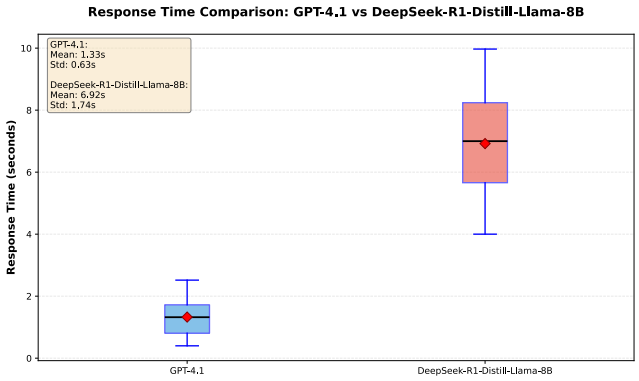

图 13：使用 GPT-4.1 和开源 Deepseek-R1-Distill-Llama-8B 时响应时间的箱线图。GPT-4.1 的延迟显著优于本地模型，表明其作为访谈聊天机器人底层模型时具有显著的性能优势。

使用 *Deepseek-R1-Distill-Llama-8B* [2] 在 *Nvidia A600* 云端运行时的均值响应时间超过 **520%** 的使用 *GPT-4.1* API 时的均值响应时间，因此在访谈期间会造成过长的延迟，尤其是在参与者数量庞大时。图 13 和 表 6 展示了这两种设置之间的响应时间差距。

表 6：响应时间对比

模型	最小值（秒）	最大值（秒）	均值（秒）	标准差（秒）
GPT-4.1	0.40	2.52	1.33	0.63
Deepseek-R1-Distill-Llama-8B	4.00	9.97	6.92	1.74

D.2 PII抽象化与删改

根据 Rescriber [78], , 小型大语言模型版本（由*Llama3-8B*驱动）和较大型模型版本（由*GPT-4o*驱动）都显著帮助用户减少了不必要的个人信息披露。不过，Llama 3-8B 的召回率（0.63）远低于 GPT-4o（0.88），因此不适合用于提供全面保护。平均而言，*GPT-4o* 版本实现了更大幅度的披露减少（总计 8.0，其中 7.4 次替换、0.6 次抽象化），相比之下，**Llama3-8B** 版本为 4.1（其中 2.6 次替换、1.5 次抽象化）。

D.3 自动 RQIEvaluation

LLM评判器是一种扩展评估规模的工具，因为人工编码员对数百或数千条回答进行人工编码极其耗时、成本高昂，而且不同评审者之间的评估结果也不一致。像*GPT-5-thinking* 这样一致性高的模型，使我们能够将人类水平的评估质量扩展到大型数据集，在忠实遵循人工制定的标准的同时，实现可扩展、可复现且经济高效的评估。我们使用*GPT-5-thinking* 和开源模型 *Deepseek-R1-Distill-Llama-8B* 开展实验，并使用RQI指标评估数据质量。*GPT-5-thinking* 的推理级别设为**低**，两个模型使用相同的提示词，对同一个包含30个问答对的问答批次进行评估。

[2] https://huggingface.co/deepseek-ai/Deepseek-R1-Distill-Llama-8B

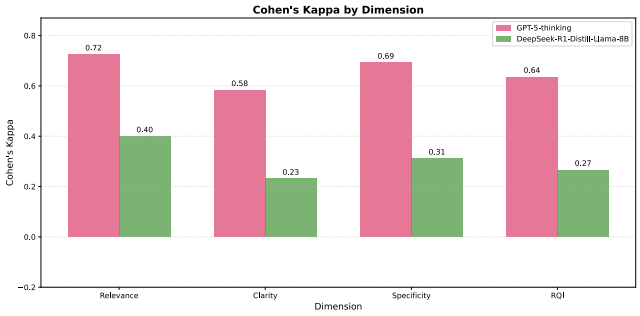

图 14：闭源模型（GPT-5-thinking）与开源模型（DeepSeek-R1-Distill-Llama-8B）之间 Cohen's κ 的比较。在理解人类偏好并与之保持一致方面，可以观察到明显的性能差距。

我们通过计算Cohen's κ，将模型评估得出的回答质量指数（RQI）与人工标注金标准进行比较。根据图14，*GPT-5-thinking* 显然比*Deepseek-R1-Distill-Llama-8B* 更适合用作RQI评估器，因为它在RQI的三个维度——相关性、清晰度和具体性——上都与人工编码员表现出显著更高的一致性。在图中，*GPT-5-thinking* 在RQI评估中达到较高的Cohen's κ 分数，而*Deepseek-R1-Distill-Llama-8B* 的表现则处于较低或接近随机的水平，这表明其判断与人工真实值的偏差大得多。总体而言，该图表明，GPT-5-thinking与人类的评估模式更加贴合，且表现更为稳定，因此是生成可靠、符合人工标准的RQI评估结果的显著更优选择。

E 原稿与删改后回答示例对（个人身份信息已更改以实现匿名化）

E.1 示例：TIME 个人身份信息删改

- **原稿：** “我于 1999 年开始工作，这算是一场非正式的访谈，当时我需要完成多份不同的文书并取得任职资格”
- **删改后：** “我于 [Time2] 开始工作，这算是一场非正式的访谈，当时我需要完成多份不同的文书并取得任职资格”
- **个人身份信息类型：** 时间

E.2 示例：地理位置类个人身份信息删隐

- **原稿：** “我是在俄亥俄州获得执业资格的律师。我在洛杉矶接受了资深律师的当面访谈。” • **编辑后：** “我是在 [地理位置1]获得执业资格的律师。我在 [地理位置2]接受了资深律师的当面访谈。”
- **个人身份信息类型：** 地理位置

E.3 示例：时间个人身份信息抽象化

- **原稿：** “我于1979年开始攻读最初的学位，并于2012年完成了最后一个研究生学位”
- **编辑后：** “我在更早的一年开始攻读最初的学位，并在最近的一年完成了最后一个研究生学位”
- **个人身份信息类型：** 时间

E.4 示例：地理位置个人身份信息抽象化

- **原稿：** “这是一家位于波士顿的科技公司”
- **编辑后：** “这是一家位于某个大城市的科技公司”
- **个人身份信息类型：** 地理位置