Orality：用语音外化和澄清思维的语义画布

Wengxi Li*创意媒体学院
香港城市大学 香港，中国
w
engxili-c@my.cityu.edu.hk

Jingze Tian*创意媒体学院
香港城市大学 香港，中国
ji
ngztian2-c@my.cityu.edu.hk

Can Liu†创意媒体学院
香港城市大学 香港，中国
canliu@cityu.edu.hk

arXiv:2603.02544v1 [cs.HC] 3 Mar 2026

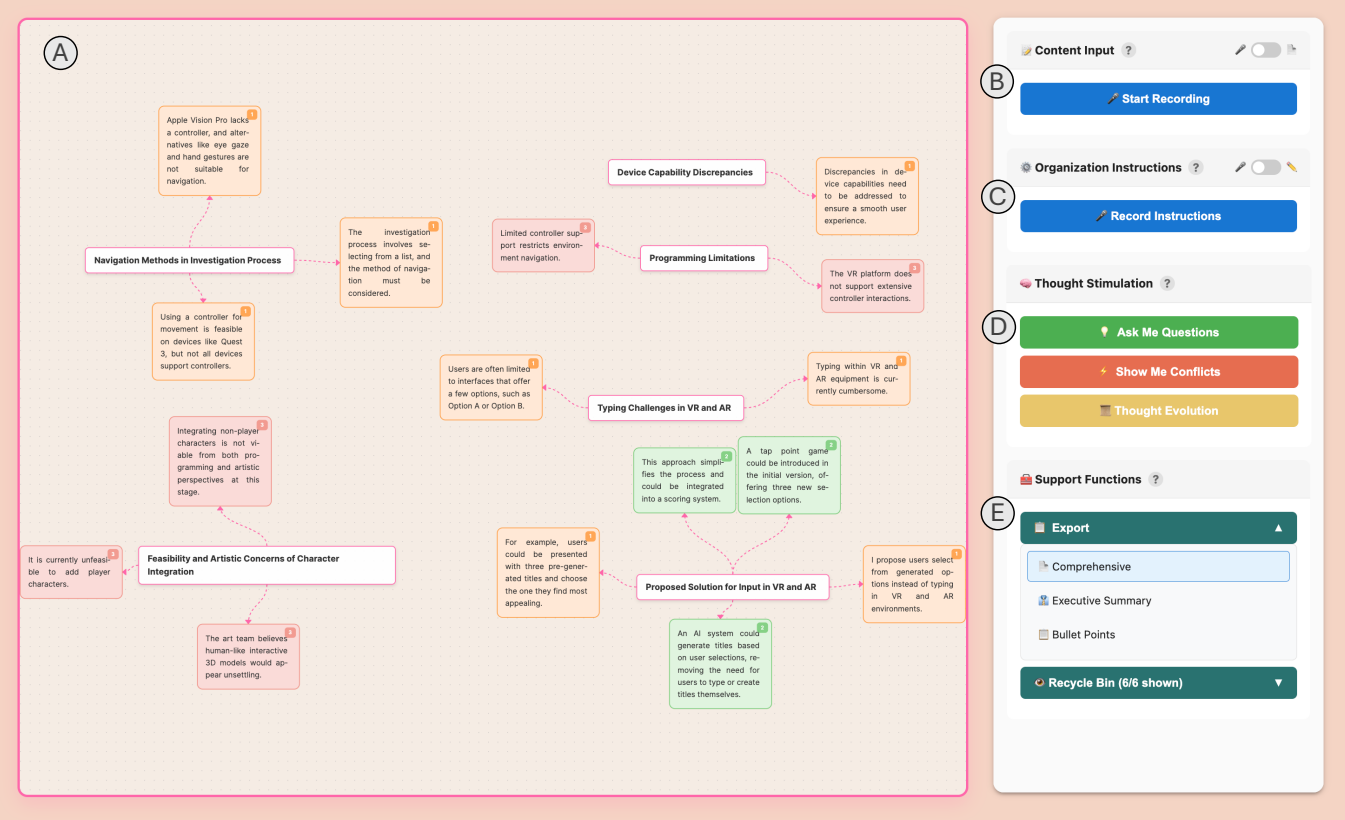

图1：Orality 的主界面与功能，展示了从语音中提取的示例内容节点和主题。通过语音输入控件（B），用户可以将其思维迭代地外化到节点链接语义画布（A）上。用户可以通过口头给出指令来调整结构和布局（C）。此外还有思维激发功能（D），可通过“向我提问”生成思维引导的问题节点，通过“显示冲突”在相互冲突的节点之间绘制链接，或通过“思维演化”浏览思维演变历程。用户还可以导出所选内容的备忘录（E）。

摘要

人们会通过大声表达来外化思维，以帮助澄清和组织思绪。尽管语音转文本可以捕捉这些想法，但由于言语不流畅、重复以及潜在的组织混乱，转录文本可能难以阅读和理解。为了支持通过语言表达进行思考，我们引入了 Orality。它从口语内容中提取关键信息，利用大语言模型进行语义分析，并在交互式画布中形成节点链接图。用户无需阅读和处理转录文本，而是可以操作节点簇，并通过口头发出指令，以其他方式重新提取和组织内容

*两位作者对本研究的贡献相同。
†通讯作者

本作品依据知识共享署名 4.0 国际许可协议授权。*CHI '26*，西班牙巴塞罗那
© 2026 版权所有者/作者持有版权。ACM 国际标准书号
979-8-4007-2278-3/2026/04 https://doi.org/10.1145/3772318.3791713

。它还提供 AI 生成的启发式问题以及逻辑冲突检测功能。我们开展了一项包含十二名参与者的实验室研究，将 Orality 与通过语音和 ChatGPT 进行交互的方式进行了比较。我们发现，Orality 能更好地支持用户澄清和发展自己的想法。研究结果还揭示了图形化和对话式思维澄清工具各自的功用，并提出了相应的设计启示。

CCS 概念
• **以人为中心的计算 → 交互系统与工具；交互技术**。

关键词
语音转文本、人机交互、大语言模型、思维澄清、思维工具

ACM参考格式：
Wengxi Li、Jingze Tian 和 Can Liu。2026。Orality：用语音外化和澄清思维的语义画布。载于 *2026*年*CHI*计算系统中的人为因素会议论文集（*CHI '26*），*2026*年*4*月*13*日至*17*日，西班牙巴塞罗那。ACM，纽约州纽约，美国，共29页。https://doi.org/10.1145/3772318.3791713

1 引言
从提出研究问题到制定商业战略，复杂想法的形成过程从根本上取决于我们外化思维的能力。我们很少在真空中思考；相反，我们会将想法转移到草图、笔记和图表等外部表征上，与自己的想法展开动态对话 [20]。延展认知这一概念认为，外部人工制品会成为我们认知过程不可或缺的一部分，这凸显了思考工具的重要性 [9]。这些工具从实体白板到数字思维导图软件不一而足，各自都能帮助我们稳定转瞬即逝的想法，使其得以审视、组织和完善。

在各种思想外化的方法中，语音是一种独特而强大的渠道。它是一种自然且快速的媒介，能够表达内心独白的流动，无须承受打字或绘图所带来的机械和身体上的阻力 [34]。心理学研究早已表明，将思维过程说出来——即“出声思维”方法——有助于人们发现逻辑中的漏洞，并将抽象概念转化为具体表达，从而改善问题解决能力 [7, 11]。口语能够捕捉自发思维特有的结构：思维通常以彼此关联的想法“簇”涌现，随后又会不连续地“跳转”到新主题 [39]。然而，语音作为一种潜在的外化媒介，也面临一些挑战。其输出本质上是线性的、杂乱无章且缺乏结构的，因此难以回顾、重新组织或综合成一个连贯的整体 [2, 5]。如今，大多数操作平台都提供语音转文本工具；近期推出的转录工具（如 otter.ai、Zoom 和手机上的笔记应用）也能生成摘要、行动事项和主题列表，以便概览内容。然而，这些输出仍缺乏清晰度或灵活性，难以作为思维媒介得到有效利用 [3]。

大语言模型（LLM）的最新进展为应对这些挑战带来了新的机遇。人们越来越多地将生成式人工智能用作思考与反思的伙伴，而不仅仅是信息检索工具 [15, 45]。用户可以与人工智能探讨复杂的个人决策或长期目标，通过对话从多个视角审视问题，并逐步厘清自己模糊的想法 [10, 24]。结合语音输入后，这些互动会变得更加自然、更具对话感，从而降低将想法外化的门槛 [25]。然而，当前基于人工智能的对话界面设计限制了其作为有效思维工具的潜力。语音版 ChatGPT 等界面受限于线性的聊天形式。随着用户不断发言、对话逐渐增多，按顺序排列的文本记录会变得难以回顾和综合，人工智能也可能难以保持上下文连贯，常常生成冗长的输出，反而掩盖而非澄清用户的核心想法 [17]。这种线性结构并不适合呈现想法之间如网络般交织的联系。尽管已有研究探索了人工智能生成的思维导图，但这类导图仍缺乏捕捉思维过程发展细节所需的灵活性与质量 [16, 36]。一些研究探索了将语音与画布结合，以灵活地外化和组织想法。但这类系统主要用于支持小组头脑风暴，或会议中的实时协作式对话映射， [6, 55] 而非支持个人的反思性思考。

在本文中，我们旨在设计并开发一种新颖的界面，以支持借助语音进行思想外化，并帮助思维进一步发展。主要目标是辅助将尚不清晰或复杂想法加以外化、结构化并激发的过程。为奠定我们的设计基础，我们开展了一项形成性研究，以了解用户在使用聊天机器人界面组织其言语化思维时所遇到的困难。我们发现，用户难以应对其非线性思维与类似大纲的线性结构之间的不匹配，并表示需要更显著、更灵活的方式来查看和操作思维结构。

基于这些发现，我们提出了 Orality，这是一个新颖的系统，它将语音优先的工作流程与由人工智能组织的语义画布相结合，以支持思维澄清的完整周期，帮助用户将模糊的初始想法转化为具体、结构良好的概念。我们通过支持在语义画布上对提炼后的口语内容进行迭代优化与空间重组来实现这一点。Orality 的设计基于一个四层概念框架，该框架将 Pirolli 和 Card 意义建构模型适配到个人思维澄清的情境中 [29]。我们用由语音输入支持的思维外化层 取代信息觅食循环，并将意义建构循环作为由画布上动态的、基于节点的数据表示所支持的结构化与图式化层 。这一框架进一步由拓展与深化层 加以增强，该层利用人工智能驱动的功能帮助用户识别缺口与冲突，以及最终的反思与呈现层 ，旨在支持元认知以及阶段性或最终输出的综合。这种整合方法旨在将线性的口语化思维流转化为图形化、可塑且具反思性的思维工作空间。

我们对 12 位参与者开展了实验室用户研究，将 Orality 的功能、有用性、工作流程和对思考过程的支持与基线进行比较。基线是一个基于聊天的大语言模型界面，使用自定义提示词生成大纲和概念图。研究结果表明，Orality 在提供

思维澄清支持方面优于基线。我们的定性研究结果揭示了语音驱动、可塑的画布式人工智能如何相较于基于聊天的人工智能，在发散思考与深入思考、主动意义建构和思想发展方面提供不同的思考支持。总之，我们作出以下贡献：

- Orality 的迭代式设计以适用于自我思想澄清的理论框架改编版为基础，并通过一项形成性研究发掘了支持借助迭代式言语化澄清思维的设计机会。
- 用于支持反思性意义建构的研究产物 Orality，采用人工智能辅助的语音转文本画布，从口语化思维中提取要旨，并将其呈现为可操作的节点链接图，同时就地提出启发思考的问题。
- 一项包含十二名参与者的被试内实验室研究，比较了多模态画布人工智能界面与基于聊天的人工智能，揭示了前者在支持主动思考和深入思考方面的优势。

2 相关工作

2.1 多种模态下的思想外化

人类思维是一种主动过程，常常延伸到大脑之外，进入周围环境。延展认知理论指出，我们会将外部人工制品——从简单的纸笔到数字设备——用作认知系统不可或缺的组成部分 [9]。第一步是思想外化：将转瞬即逝、存在于内心的想法转化为稳定的外部形式，这是一项耗费心力的任务 [42]。根据双重加工理论，认知过程分为两种：一种是无意识、快速、自动且容量大的过程（称为系统1），另一种则是有意识、缓慢且深思熟虑的过程（称为系统2） [12, 19]。早期想法的直接流露符合快速、直觉的系统1加工，且往往基于抽象的要旨；但记录这些想法的行为可能会带来较高的认知负荷，并要求采用缓慢、分析性的系统2方式，才能形成逐字记录 [32]。因此，选择何种模态和工具，对于减少这种认知摩擦至关重要。

人机交互研究已探索了多种采用不同模态的工具来支持外化，例如书写、草图绘制、手势和语音。通过键盘和文本编辑器以书写方式进行思想外化是最常见的 [14, 21]。数字笔记应用、大纲编辑器以及Mindalogue [51] 和VISAR [54] 等思维导图工具为用户提供画布，以便通过书写记录想法。另一些系统则探索利用沉浸式环境将想法空间化 [49]。它们表明，空间组织和直接操作对于帮助用户发现模式、构建心智模型并发展更多想法至关重要。

除了这些基于视觉和文本的系统，近期研究也开始探索将语音作为外化思维的媒介。该领域的奠基性工作，例如 Bolt 的“Put-That-There”系统，展示了如何将语音和手势结合起来，作为控制图形界面的自然模态 [4]。Rambler [25] 支持通过口述撰写长篇文本，并使用大语言模型帮助用户整理和编辑内容。LADICA [55]和 MeetMap [6] 会从小组对话的语音中提取信息并呈现在画布上，以促进头脑风暴活动。然而，这些工具往往着重于协助生成某种输出（例如文章或代码），方法是对用户生成的内容进行后处理，或根据外部资源提出想法。针对捕捉、澄清和梳理模糊思维的工具和研究仍然不足，尤其是利用语音输入模态的这类工具和研究。

我们的工作旨在通过创建一个无缝整合两种思维模式的界面，支持思想外化与思想组织之间更流畅、可迭代的转换：从系统1所特有的快速、联想式外化，到系统2审慎、分析式的组织。我们认为，基于语音的语义画布能够有力地支持这一目标。我们的系统利用语音创建节点链接画布，以呈现口语内容中的语义，从而帮助用户从创作过程的最初阶段就开始外化、组织和迭代自己的想法。

2.2 用于文本意义建构的人工智能辅助画布界面

厘清思路并非简单的转录行为，而是一个主动的、建构性的意义建构过程 [48]。想法一经外化，人们便可以开始整理、关联并完善这些想法，使其形成连贯的结构。人机交互研究在构建支持这一意义建构循环的可视化工具方面历史悠久。一些视觉意义建构工具允许用户在空间中排列笔记，形成主题簇和概念图，例如用于口头内容的 TalkTraces、用于视频的 VideoMap，以及用于大语言模型生成的文本的 Sensecape 和 Graphologue [5, 18, 23, 41]。这些工具引入了自动发现关联的功能，并提供了探索思路的空间，支持意义建构循环中的图式构建环节。Luminate 和 Patchview 等系统可以基于对多个语义维度的意义建构来生成故事 [8, 40]。ConceptScope 和 ConceptEVA 等其他系统则着重帮助用户通过反复迭代来定义并发展对某个主题的概念理解 [52, 53]。这些工具为整理丰富的信息提供了强有力的支持。

尽管这些可视化技术在分析外部数据方面十分有效，但如何将其加以改造，以支持更为个人化、更为流动的思维意义建构过程，仍是一个有待探索的开放领域。这一过程有所不同，因为信息并非事先存在，而是在同一时刻被生成并加以结构化。传统的内容创作工具，如文本编辑器或大纲编辑器，虽然天然地将外化与组织结合在一起，但往往缺乏现代意义建构工具所具备的视觉与语义结构化能力。机遇在于，将流畅的、最初的出声思考行为，与动态、灵活的可视化画布的力量连接起来。

我们的工作旨在通过在思维外化与意义建构之间建立迭代循环来弥合这一差距。借助语音实现流畅而高效的输入，我们的系统让用户在各次迭代之间看到自己的想法以可操控的外部对象形式呈现，并支持通过口头指令进行高度灵活的重组。这一设计旨在降低阐述一个想法

与操控它之间的障碍，从而鼓励更加动态的言说、观看与结构化循环。该方法将人工智能辅助的可视化技术用于一项独特挑战：帮助用户理解自己正在涌现的想法。

2.3 借助元认知工具支持思考

有效的思维工具不应止于捕捉和组织信息，还应支持思考本身所涉及的认知过程。理解这类支持的一种实用框架是元认知，即监控和控制自身认知过程的能力 [13]。元认知包含两个主要组成部分：监控，即评估自身的理解和进展（例如，“这个论证连贯吗？”）；控制，即引导认知活动以实现目标（例如，“继续之前，我应该进一步探索这个想法”） [27]。如何设计能够明确支持元认知的工具，仍是人机交互领域的一项核心挑战。

现代计算工具，尤其是生成式人工智能（GenAI）的出现，对用户提出了新的、更高的元认知要求。正如 Tankelevitch 等人所指出的，与生成式人工智能系统交互需要投入大量元认知努力：用户必须持续进行规划，例如拟定有效的提示词；进行监控，例如评估生成内容的准确性、相关性和质量；并进行控制，例如决定何时信任人工智能、何时修改其输出，以及如何将其整合到更大的工作流程中 [45]。人机交互领域正积极探索在人工智能支持的工作新时代，如何理解、保护并增强人类认知 [44]。一种方法是提供认知支架，由系统提供结构和指导，帮助用户管理自己的过程。例如，一些系统会使用元认知提示，鼓励用户在使用生成式人工智能进行搜索时开展更多批判性思维 [38]。其他研究表明，在学习环境中，支架有助于提升问题解决能力和元认知技能 [47]。另一种方法是提高系统透明度，帮助用户建立对工具能力和局限更准确的心智模型，进而提升他们监控和控制交互的能力 [22]。这些系统旨在让用户更加意识到自己的思维过程，并提供有效引导这一过程的方法。

我们的工作通过设计一种系统，为思维形成的早期阶段降低元认知负荷，从而为这一研究方向作出贡献。一些研究者提出，人工智能应当扮演“外部辅助者”的角色，成为帮助培养高阶思维技能的伙伴 [50]。我们的系统与这一观点一致，通过提供思维激发功能来支持而非主导思维过程。此外，语义画布的自动生成会呈现外化思维中的结构和关系，为监控过程提供直接的视觉反馈。

3 形成性研究

我们开展了一项形成性研究，以揭示在澄清言语化思维时提供额外交互与视觉支持所面临的设计挑战和机遇。研究结果为我们所提出原型的功能设计提供了依据。

3.1 任务、参与者与流程

本研究要求参与者体验这样一套工作流程：使用语音转文本输入创建言语化内容，并通过提示大语言模型对其进行整理。任务开始前，参与者被要求选择一个与近期事件相关的开放式主题，该事件需要进行思维澄清（例如，梳理一个令人困惑的项目，或深入反思某个问题）。这样可以确保内容源自他们的近期记忆，并切合他们对思维澄清的实际需求。

我们从一家研究机构招募了八名参与者，其中五名自我报告为女性，三名为男性。所有参与者都有使用语音转文本工具的经验，并且日常会使用大语言模型完成任务。每位参与者均因参与研究获得 15 美元报酬。每次研究会话约持续一小时。

在研究的前半部分，参与者自行提出一个主题，并获得至少 30 分钟的时间来口述和整理思路，直到形成最完整、最理想的版本。在任务过程中，他们使用 Google 文档的语音转文本功能输入内容，并为大语言模型产品 DeepSeek[1] 编写提示词，以根据需要整理转录文本。

在下半部分，参与者接受了访谈，反思自己的体验，重点讨论在口头表达和基于大语言模型的组织过程中对额外支持的需求。整个会话均被录音、转录，并通过主题分析进行分析，以识别设计挑战与机遇。

3.2 研究结果与讨论

本形成性研究探讨了参与者如何针对各种自行提出的任务外化并组织自己的想法，包括项目规划（P1、P2、P4）、创意写作（P3、P5、P6）和问题解决（P7、P8）。他们口头表达的内容通常捕捉了参与者内部思维的自由流动，包括零散的想法、未成形的言语以及突发的思绪火花，类似于未经筛选的出声思维过程。附录 B 展示了一个具有代表性的转录文本示例。通过对其过程的分析，我们确定了四个关键主题，揭示了重要的挑战与设计机遇。

3.2.1 对非线性思维表征的需求。报告中最根本的挑战，是思维的非线性本质与口述的线性本质之间固有的冲突。这种线性的原始思维流产生了一串文本，其中交织的内容结构难以阅读、理解和管理。参与者将自己的初始过程描述为说出“想到什么就说什么”（P2），结果形成了难以管理的漫无边际的内容。P1 明确表示，其“漫无边际的内容无法轻易删除或按层级分类”。这促使他们明确希望采用不同的表征方式。正如 P1 所建议的，理想的界面应允许他们查看大语言模型的反馈，以“理解正在讨论的主题以及任何遗漏的要点”，这表明需要主题建模和易于理解的概览，而不是一团转录文本。

3.2.2 对灵活、层级化且由用户定义的结构的需求。参与者始终会尝试为自己的

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

思维施加结构，无论是在将思维 verbalize（用语言表达）之前还是过程中。当他们先对内容进行高层规划，再深入具体细节时，就会出现这种情况。例如，P2 明确地先概述了演讲的整体支架："我打算围绕这个主题主要讲：要做什么、研究问题、如何设计实验、时间线，……"，之后才充实内容细节。同样，P5 首先规划了叙事的"蓝图"——背景、人物及其角色——以确保故事连贯，然后再展开实际内容。这些结构高度依赖任务，并由用户主导。P1 用三部分结构（过往进展、当前状态、下一步）组织进行中的项目，而 P4 则通过"沿着时间线填入要做的事情"来规划活动。此外，参与者还需要层级组织方面的支持。P1 提出的"整体安排"需求，以及 P2 使用自行填写的多级大纲，都表明他们需要能够同时呈现思维结构的高层视图和详细视图的工具。

3.2.3 支持迭代思考过程的必要性。 研究表明，思维澄清并非线性的一次性过程，而是高度迭代的过程。P1和P5均指出，他们的思考通过"复述"或重新阐述观点而得以深化，P5表示，口头陈述过程中的自我反思带来了"对主题更深入的反思和想法"。这一迭代过程在基于聊天的界面中面临重大挑战。参与者难以在不产生冗余的情况下，将新输入与现有内容合并。例如，P1描述道，他们"试图以连贯的方式解释一些内容，但效果不佳（重复很多），因此我决定在全部完成后将它们统一整理"。同样，P5发现大语言模型合并新段落的尝试并不有效，于是他们"被动接受了零散的组织方式"。这表明，迫切需要为迭代且可控的内容演进提供更好的支持。

3.2.4 对非侵入式大语言模型辅助与用户自主权的需求。参与者一致认可大语言模型从其语音中提取要点的能力，例如"它提取了我在随口一提中说到、后来却忘了的要点"（P2），但他们在使用其辅助功能时也遇到了不少挑战。一个反复出现的主题是：有用的自动化与用户控制权丧失之间存在权衡。参与者表示，大语言模型可能会误解他们的意图，导致生成"毫无意义、重复的内容"（P1、P2），或过度解读他们的想法（P3、P4、P5）。对于创意任务，P5认为大语言模型补充的内容"太过刻板"，甚至可能"让他们偏离原本的思路"。P1阐述了人工智能理想的角色：他们希望大语言模型能够提供反馈、指出遗漏之处，并促使他们补充更多细节。这表明，人们可能需要一种智能辅助，既能提供有价值的建议，又不会凌驾于用户的控制权或创作方向之上。

3.3 设计目标

基于我们形成性研究的发现以及相关理论的支持，我们综合提炼出四项设计目标，用以指导系统的开发。这些目标旨在直接应对用户语音驱动的思路澄清过程中所识别出的关键挑战与机遇。

DG1：支持口语思维的语义检视。为应对线性语音与非线性思维之间的冲突，我们旨在将顺序语音自动转换为二维空间表示。这是为了借助外部表征支持思维 [20]，从而卸下仅在头脑中保持并发展思维结构所带来的记忆与计算负荷 [35]。通过将口语数据映射到视觉空间，我们希望帮助用户以视觉方式检视语义关系及其论证的"形态"。

DG2：提供灵活的表征方式，以便通过语音表达和重组结构。 参与者需要能够随思维变化而演进的动态结构。以往研究也表明，基于语音的文本创作往往会外化意识流，从而形成不断演进的结构 [25]。因此，我们希望用户能够通过灵活的语音指令定义、应用并反复重构其组织框架，使结构随着用户内部心智模型的演进，从模糊逐渐变得清晰。

DG3：呈现思维演化的迭代变化。思维会在反复修订中不断深化，因此我们旨在使这一演进过程清晰可见且便于浏览。社会科学研究早已证明，帮助专业人士进行"行动中反思"至关重要 [37]；这种帮助可以让他们暂停、重新审视视角，并质疑自己的思维过程。为了支持元认知反思 [13]，，系统应允许用户查看并比较其思维结构的历史版本，以评估进展并完善逻辑。

DG4：在保持用户自主权的同时提供智能内容支持。为应对辅助与用户意图之间的权衡，我们旨在按需提供支持，增强用户能力而非喧宾夺主。混合主动交互领域的既有研究 [1] 表明，系统应提供价值（例如，识别空缺或冲突），同时不削弱用户的主动性。我们的系统旨在为思考提供支架，而非控制内容或打断创作心流。

4 设计 ORALITY——人工智能辅助思路澄清工具

基于形成性研究的发现和设计目标，我们开发了 Orality。这是一个人工智能增强型语义画布，支持用户通过语音反复外化、完善和重组想法，从而澄清思路（图1）。本节其余内容安排如下：首先，我们将在第4.1节介绍设计背后的理论；然后，第4.2节将介绍 Orality 的主要功能；第4.3节将通过一个示例场景展示如何将 Orality 用于研究项目；最后，第4.4节将介绍实现细节。

4.1 理论基础——自我思路澄清框架

我们提出了一个用于自我思想澄清的概念框架，为 Orality 的设计奠定基础。该框架是对 Pirolli 和 Card 提出的**意义建构模型**的一种实例化（图14

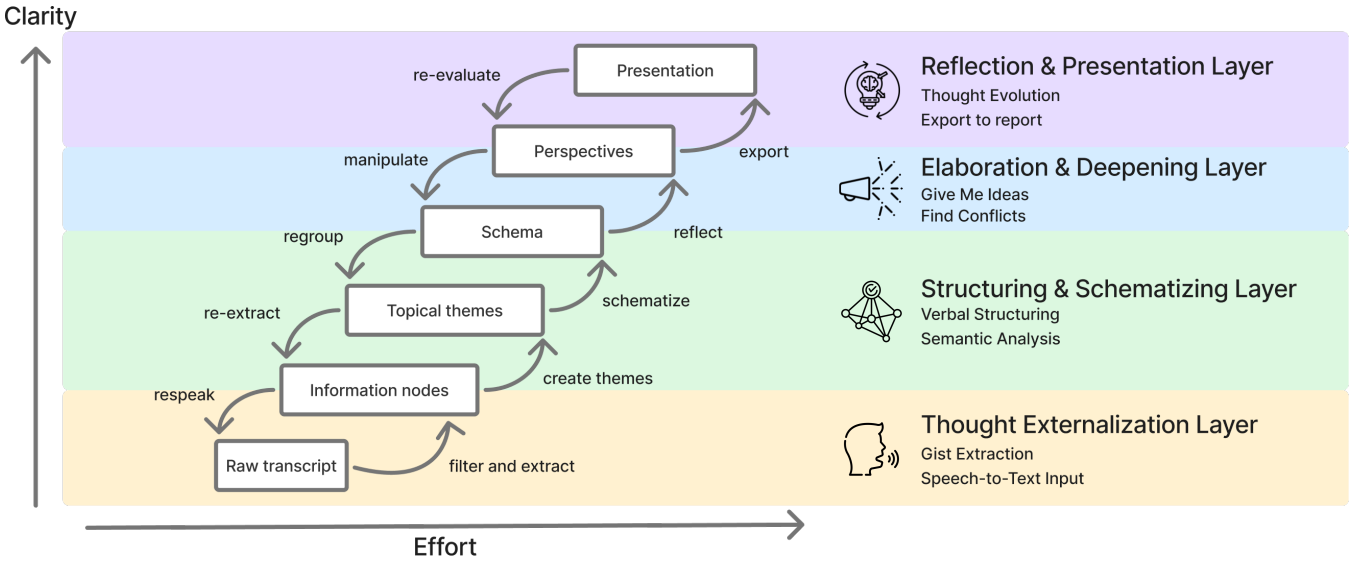

图2：我们的自我思想澄清过程概念框架，改编自 Pirolli 和 Card 的意义建构模型 [29]，，其中的各个阶段对应于 Orality 中设计的四层系统支持。每一层都列出了相关的系统功能。

见附录 A） [29]，。该模型最初旨在描述分析人员如何与大型数据集交互以进行意义建构。我们的形成性研究发现，用户面临的挑战各不相同，但都与这一意义建构过程相契合。首先，静态转录文本无法满足用户对灵活层级结构的需求，而这类结构应随着用户理解的深入而不断演变。其次，这一过程具有迭代性，用户往往需要补充细节或重新表述想法以使其清晰，但他们难以跟踪这些变化，也难以管理可能偏离其意图或引入冗余的人工智能辅助。这些发现表明，思维澄清本质上涉及对自身模糊思绪进行组织和迭代，归根结底是一项针对外化思维的个人意义建构任务。

原始意义建构模型包含一个双循环过程（附录 A 中的图 14）。第一个信息觅食循环 涉及收集原始数据，以构建一组基础的相关信息（“鞋盒”）。随后，第二个意义建构循环 以这些信息为基础，反复将其组织成连贯的图式，并据此检验假设。在我们的研究情境中，用户要理解的是自己的内部数据——也就是自己的思想。因此，我们对这一框架进行了调整，将意义建构过程的核心阶段映射到表达和组织个人思想这一相应过程中。我们将信息觅食循环 替换为思想外化过程。我们还对意义建构循环 作了轻微调整，将其目标改为通过反思生成视角，而不是依据图式化证据生成假设。数据意义建构过程通过付出努力来增强结构 ，而我们的过程旨在提升清晰度。

图 2 展示了我们的概念框架，阐释了自我思维澄清各步骤中涉及的行动和材料。我们的系统为这一过程的每一步提供四层功能支持。我们将在下文详细介绍系统的各项功能。

4.2 关键功能

在本部分中，我们介绍思维澄清过程各阶段的关键功能，这些功能已在示例场景中得到展示。

4.2.1 思维外化层。 该层借鉴信息觅食循环的概念，能够捕捉用户的语音，并创建用户想法的视觉呈现。该层旨在为思想的初步外化提供辅助，以应对工作记忆限制以及非线性思维与线性语音不匹配所带来的挑战。为此，Orality 将迭代式语音听写与动态语义空间相结合，在想法被说出时自动对其进行过滤、提取并聚成簇（图2底部）。

Orality 以迭代式语音听写作为其主要输入方法。用户可以连续不断地、并分多轮说出自己的想法。从宏观层面直接看到已说出的内容，对于建立这些想法的初始心智模型至关重要（DG1）。Orality 通过将口述输入转化为空间化且可交互的语义内容空间来实现这一点（图1）。我们称之为“语义画布”，以表明其中的元素及其空间组织并非任意安排，而是由算法生成，用以表征一组信息内部的语义关系。这块二维画布作为用户心智模型的外化表征，使用户能够识别那些难以理解、也难以记住的模式与关系。

我们选择使用节点链接图来表示语义空间，是为了应对口头内容具有线性、缺乏结构且转瞬即逝的特性所带来的挑战。

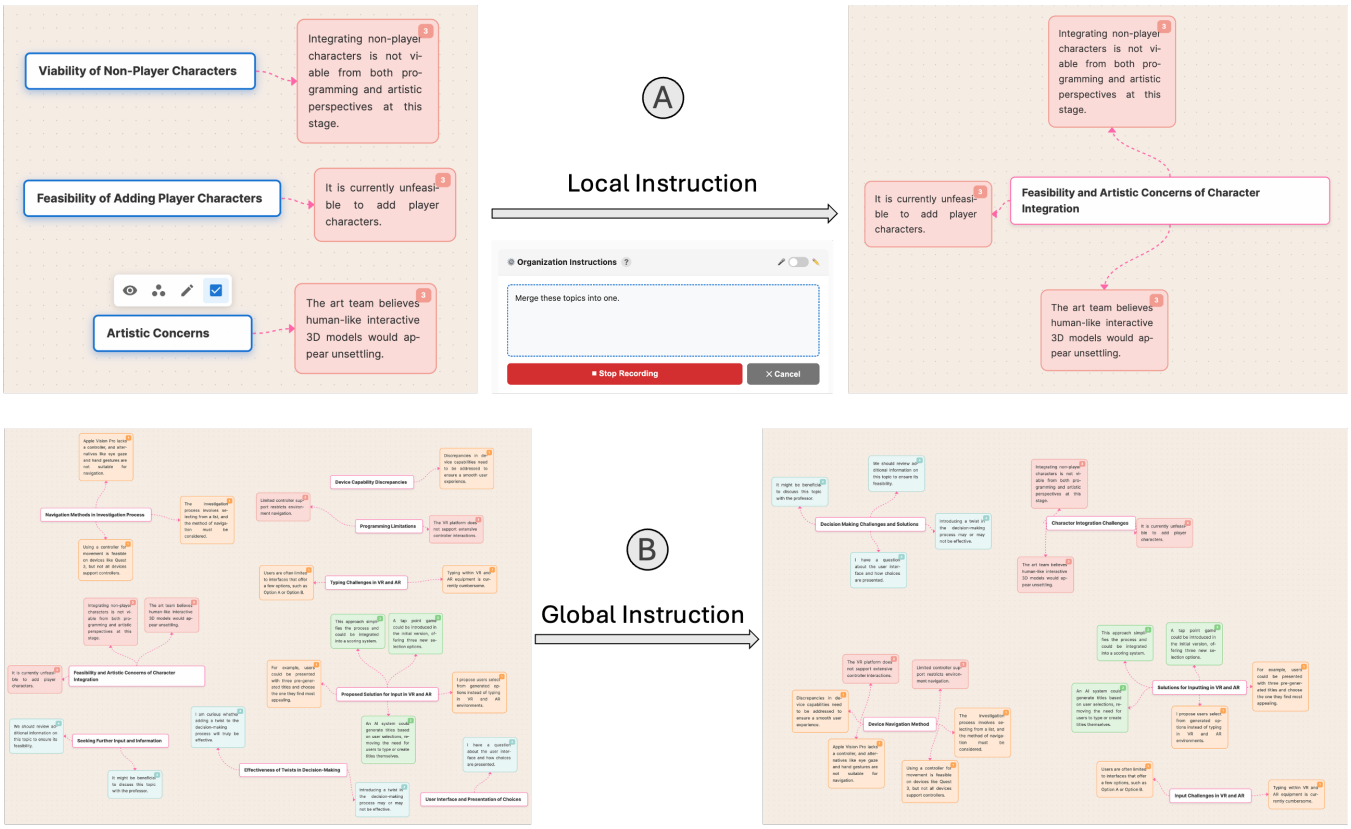

图 3：口头重组：用户既可以指定仅适用于选定主题组的局部指令，例如合并请求（A）；也可以指定适用于所有主题组的全局指令，例如特定的新结构（B）。

概念图研究表明，这类表示通过使想法之间的语义关系变得明确、持久且可灵活调整，能够降低认知负荷并支持意义建构 [28]。其非线性和层级化的表示方式让用户能够总览自己的想法并审视其中的交叉链接结构，同时持续拓展思维脉络。它还提供了一个灵活的基础，能够表示多种结构，以适应用户创建的不同主题，从旅行计划到说服性论证不等，因此适合作为通用的思维澄清工具。该层的技术细节见第4.4.3节和第4.4.4节。

4.2.2 结构化与图式化层。第二层旨在支持用户通过口述为其外化思维构建组织结构（DG2）。这对应于意义建构循环中自上而下、由假设驱动的活动：用户主动重组想法，以形成连贯的图式（图 2 下方中部）。其核心功能是口头结构化机制，用户可通过该机制输入口头指令，重新组织语义内容空间（图 3）。这一过程使用户能够高效地测试不同的组织图式，而无需进行大量手动操作。

重组指令可应用于两个控制层级。对于广泛的自上而下调整，用户可以输入一条全局指令，使其适用于空间中的所有节点（图 3 A）。对于更有针对性的调整，用户可以先选择空间中的一部分特定主题节点，然后输入一条局部指令，使其仅作用于选中的项目（图 3 B）。系统支持多种重组命令。用户可以通过口述所需的确切新主题，请求特定的新结构，例如：“将选中的内容重组为三个主题：‘设计目标’、‘用户反馈’和‘技术限制’。”或者，用户也可以在不指定主题的情况下请求新的结构，促使系统分析内容，并根据语义分析生成新的组织结构。

此外，系统还支持通过修改现有结构的指令，动态、迭代地演化图式。用户可以口头指示系统添加新主题，系统会从转录文本中识别相关子节点，并将其移至新建主题中。用户还可以选择多个主题节点并输入指令，将它们合并为一个整合后的主题；或者选择单个主题并指示系统将其拆分为多个更细致的主题。这些口头指令为用户完善思维组织提供了一种灵活、省力的方法，直接支持构建和测试心智模型的迭代过程。

4.2.3 深化与拓展层。在将思考外化并加以结构化之后，第三层旨在帮助用户

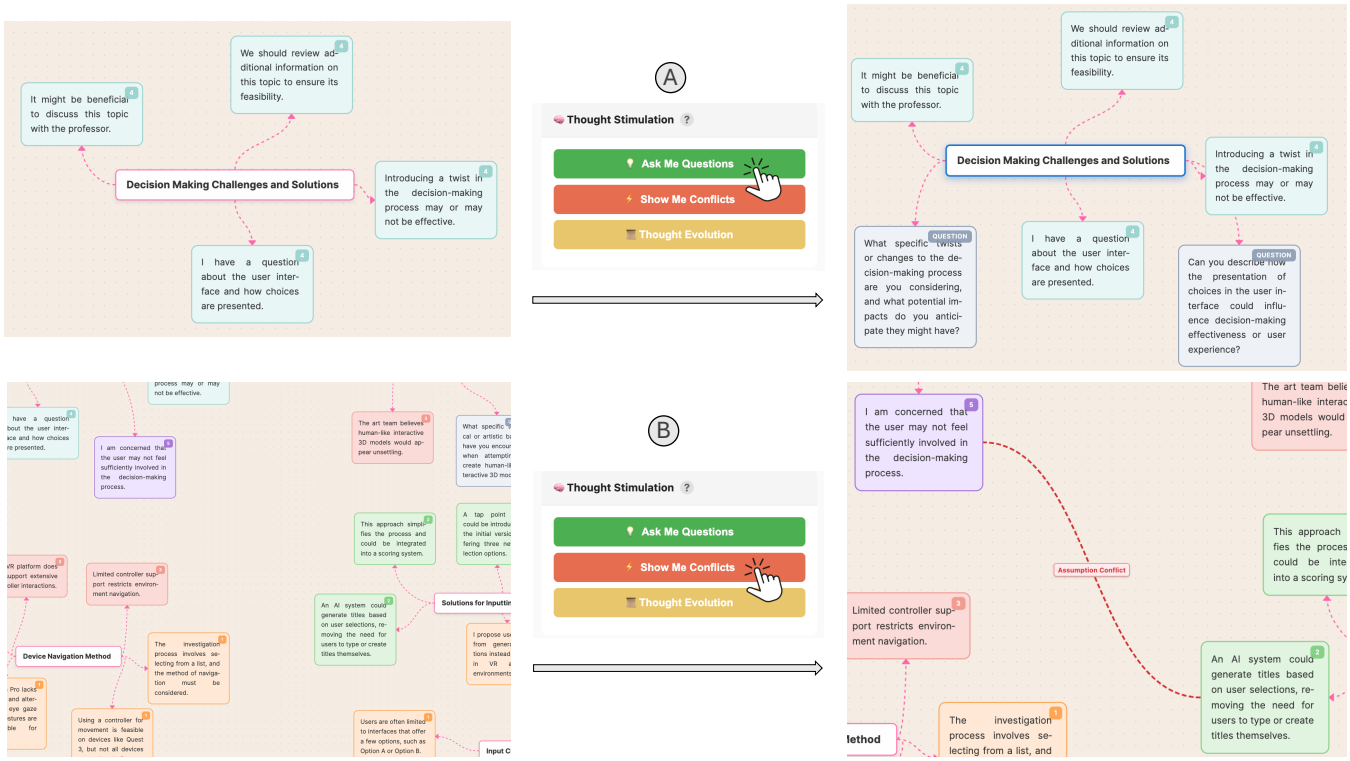

图4：原位 AI建议：A：对于现有的主题组“决策挑战与解决方案”（左），用户点击“向我提问”按钮，系统会生成两个与该主题组关联的引导性问题（右）。B：点击“显示冲突”按钮后，系统会检测节点之间的逻辑冲突。如果发现冲突，系统会在节点之间生成一条虚线，并标注该冲突（右）。

加深理解并拓展更多视角，方法是识别尚未充分发展的部分并发现潜在冲突（图 2 上方中间）。Orality 提供的功能能够找出用户当前图式中的薄弱之处和机会，从而加速意义建构循环，推动用户从组织阶段迈向更全面、更连贯的最终成果（DG4）。

侧边面板中设有一个标记为“向我提问”的按钮（图4 A）。此功能既可应用于整个画布，也可应用于选定的主题。未选定任何主题时触发该功能，系统会分析所有主题中的内容分布，并找出关联内容节点数量相对较少的主题。随后，系统会在该主题中生成作为新节点的引导性问题，以促使用户进一步输入。或者，用户也可以先选择一个或多个特定主题，再触发此功能，这样系统只会在选定的主题中生成相关问题。

“显示冲突”按钮会指示系统分析所有内容节点的语义内容，以查找逻辑不一致或相反陈述（图 4 B）。此过程旨在识别几种逻辑问题，包括直接矛盾（例如，一条陈述称某项功能“必不可少”，而另一条则称其“没有必要”）、逻辑不一致（例如，两条无法同时为真的陈述）以及策略冲突（例如，针对同一个问题提出两种互不相容的方法）。检测到冲突时，系统会在存在冲突的节点之间绘制连线，并在连线上标注冲突类型，以此将冲突可视化。

4.2.4 反思与展示。层。最后一层支持意义建构过程的最终阶段：重新评估和展示（DG3）。用户可以通过交互式时间线回顾思维过程的演变，从而增强对想法如何发展的元认知意识。此外，该层还支持最终展示阶段，将动态的思维空间转化为简洁连贯的备忘录（图 2 上方）。这些构成了意义建构过程的最后阶段，在此阶段，一组经过提炼的想法将被整理好，以便交流或用于未来行动。

系统会在用户会话期间的关键时点自动保存版本历史，例如用户口述新内容或完成对节点的操作时。这段历史以时间线的形式直观呈现，用户可以访问（图 5 A）。当用户将光标悬停在时间线上的某个版本节点上时，Orality 会展示该时刻语义内容空间的预览，并通过平滑动画在当前状态与预览状态之间切换。如果用户点击该版本，主界面便会恢复到相应的历史状态，让用户回顾工作，甚至从先前的进度继续进行。

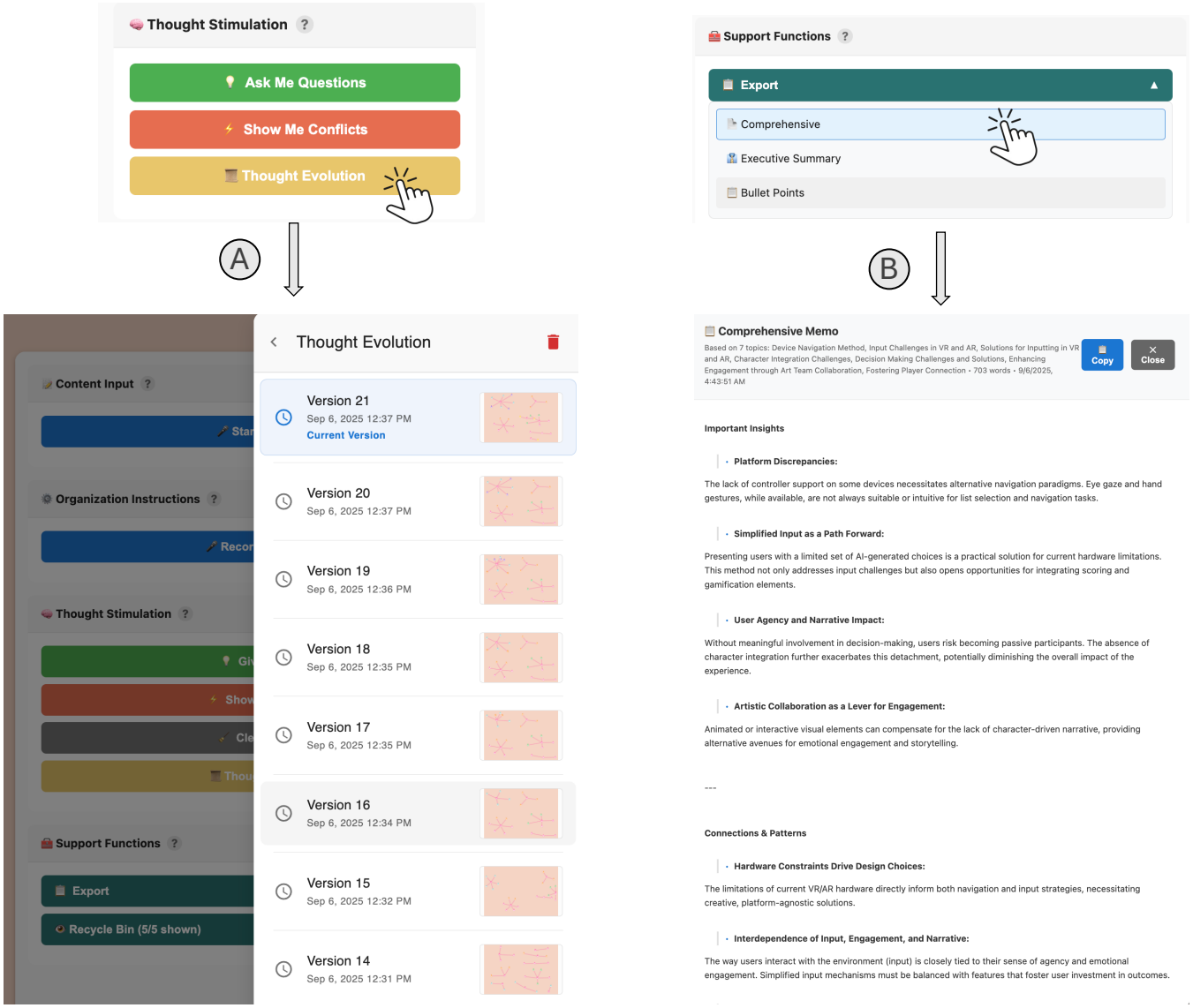

图 5：A：点击“思维演进”按钮会弹出一个时间线窗口。将鼠标悬停在时间线上，画布布局便会通过平滑的动画过渡恢复到相应阶段。B：系统提供三种“导出”格式，可针对选定的画布元素生成结构各异的报告。

“导出”功能会为用户选中的内容节点生成文本摘要（图 5 B）。用户可以根据目标从多种格式中进行选择。全面摘要会综合关键主题、指出关联，并以反思性的语气提出可能的后续步骤。执行摘要提供高层次概览，重点呈现最重要的见解和战略影响。最后，要点摘要会将想法整理成结构化列表，方便快速查阅。这些导出选项为思路澄清过程提供了可切实使用的输出，适用于记录或分享。

4.3 示例场景

本部分将介绍一个场景，展示研究生 Thomas 如何使用 Orality 整理他对一项正在进行的研究项目的想法。

内容听写与布局创建。Thomas口述了最初未经整理的想法。Orality提取出四个关键部分：“调查过程中的导航方法”“设备能力差异”“VR与AR中的打字挑战”和“VR与AR输入的拟议解决方案”，并将它们显示为节点链接图（图1）。为补充最后一个主题的细节，他选中该主题并再次口述；Orality将新内容提取到该集群中，并用不同颜色标示第二轮输入（图6）。

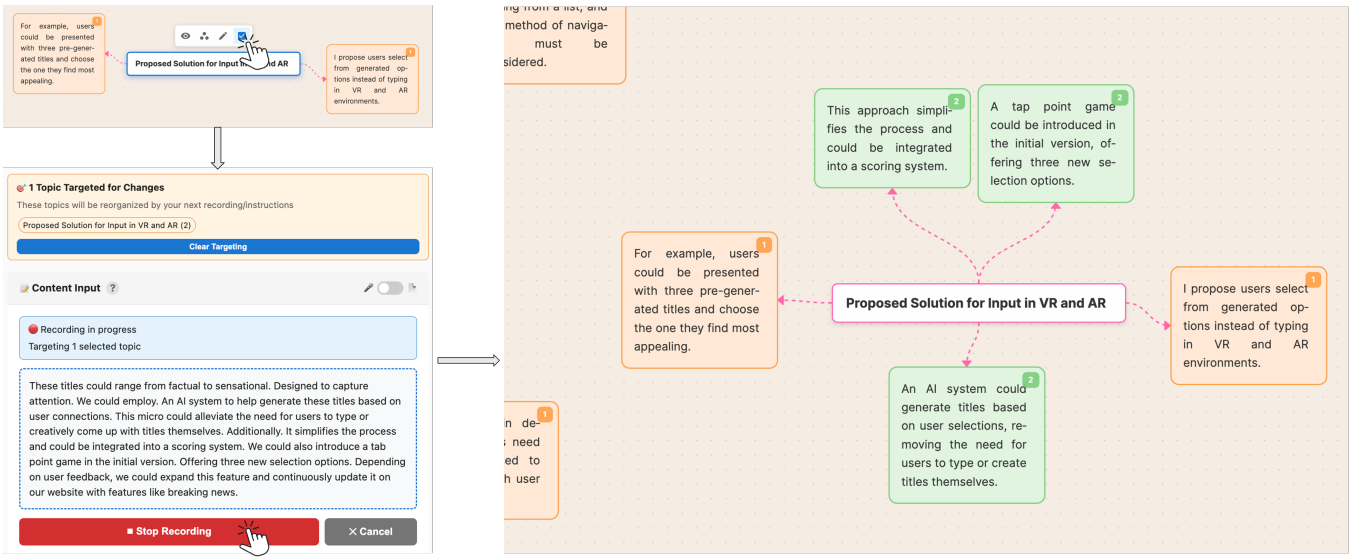

图6：迭代式言语化：通过选择主题组（“VR 和 AR 输入的拟议解决方案”）并进行口述，系统会分析口述内容，并将其作为不同颜色的新节点添加到所选主题组下。

通过重组指令切换视角。为整合相关想法，Thomas选中多个节点并口述：“将这些主题合并为一个”（图3 A）。之后，为了从全新的角度审视项目，他录制了一条全局指令，并指定了新的结构。Orality根据这些新主题重组了整个信息空间（图3 B）。

思维激发。为了进一步阐述“决策”，Thomas 对目标节点使用“向我提问”。系统生成相互关联的问题节点（图4 A），促使他进一步言语化。随后，他启用“显示冲突”，该功能揭示了两个节点之间的一条“假设冲突”边（图4 B），促使他删除有缺陷的内容。

回顾并导出想法。Thomas 使用“思维演进”浏览时间线，预览并恢复到先前的画布状态，以追溯其逻辑（图5 A）。最后，他点击“导出”，生成一份全面报告，总结关键见解和后续步骤，以便存档（图5 B）。

4.4 实现

Orality 被实现为一个 Web 应用程序，包含基于 React 的前端客户端和基于 Flask 的后端服务器，负责管理数据处理、实时通信以及与大语言模型的交互。技术细节将在以下各部分中讨论。每项功能所使用的提示词见附录 E.2。

4.4.1 前端架构. 用户界面使用 React 库开发，并采用 TypeScript 确保类型安全性和可维护性。界面的核心是语义内容空间，使用 React Flow 库（@xyflow/react）进行渲染。该库为创建基于节点的编辑器提供了强大的框架，我们用它来显示和管理主题节点与内容节点。客户端与服务器之间的通信通过 Socket.IO 处理，从而实现实时、双向、基于事件的通信。这对于实时转录等功能，以及接收服务器更新（例如新生成的想法或识别出的冲突）至关重要。UI 组件使用 Material-UI（@mui/material）库构建，该库提供了一致且响应式的设计系统。

4.4.2 后端架构。 后端服务器使用 Python 构建，采用 Flask Web 框架和 Flask-SocketIO 扩展实现实时通信。为了实时转录用户的口述内容，服务器接入 AssemblyAI API，流式传输音频数据并接收返回的转录文本。核心自然语言处理与生成任务通过调用 OpenAI 的 GPT-5 API 完成，包括主题提取、问题生成、冲突检测和报告生成。

4.4.3 语义分块与实体提取。 如图 7 A 所示，Orality 通过使用大语言模型执行语义分块提取，来处理包含言语不流畅、重复和非线性思维模式的非结构化连续语音。为此，我们的提示首先将关键信息提取为短句（内容节点），而非关键词或短语，从而使节点包含足够的信息，同时不会丢失过多细节。共享同一“主题”（主题牵引器）的句子被分组在一起，代表说话者陈述或观点。为在信息密度与视觉清晰度之间保持功能性平衡，我们将大语言模型限制为每个发言轮次的每个主题生成一到三个实体。该范围是我们根据经验确定的；若允许每个主题在每一轮拥有超过三个实体，会导致节点信息过度碎片化，并容易造成重叠，而少于三个则常常导致过度抽象，无法保证信息的原子性。当用户在后续轮次中继续发言时，系统会判断新的输入是扩展现有主题

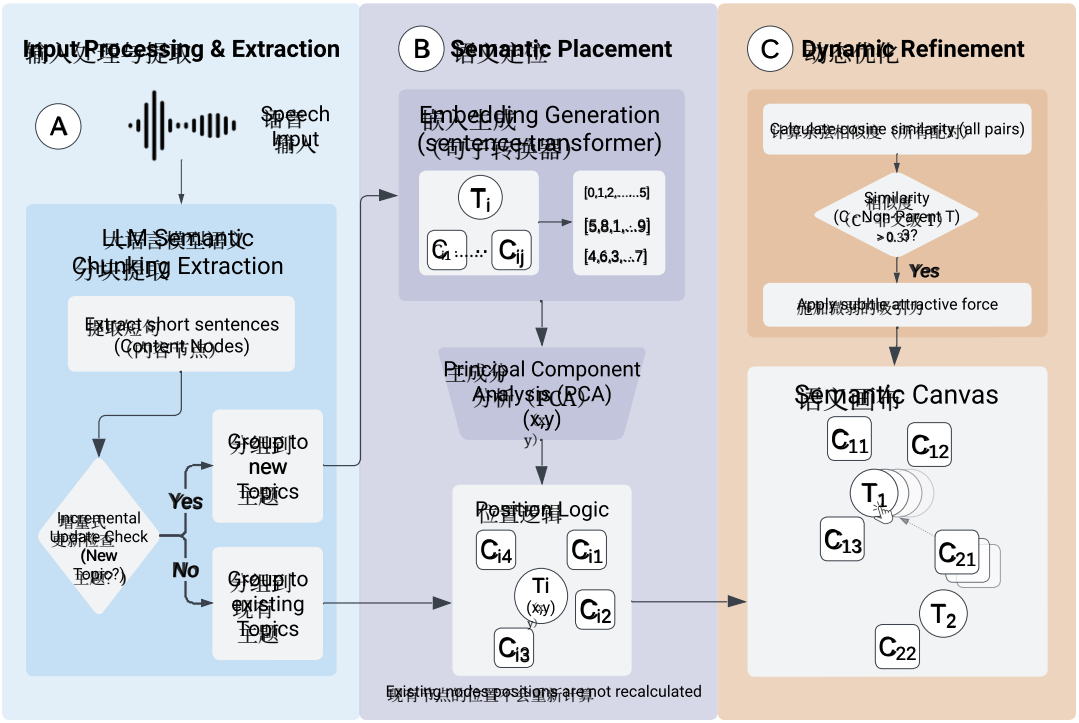

图 7：Orality 中语音到画布处理的系统架构。该流程包含三个部分：（A）输入处理从语音中提取内容节点，并将其分组成主题；（B）语义定位使用基于 PCA 的降维方法对节点进行空间定位，同时保留现有节点的位置；（C）动态优化在语义相似的节点之间施加吸引力，以保持空间连贯性。

还是引入新的主题。这使得画布能够增量更新，从而支持口语化思维的非线性特征，即用户可能在多个轮次中对先前的观点加以阐述。

4.4.4 两阶段语义布局。 节点的空间排列由两阶段语义布局计算决定。第一阶段 语义定位 在大语言模型从语音中提取主题和内容后，于后端进行。如图 7B 所示，服务器首先使用 sentence-transformers 库为每个主题节点和内容节点生成高维向量嵌入。为了将这些高维关系转换为二维布局，我们使用 scikit-learn 进行主成分分析（ PCA ）。此步骤将嵌入向量的维度降至两个主成分，随后将其用作每个主题节点的初始 (x, y) 坐标。这确保语义内容相似的主题最初彼此靠近。

我们选择 PCA，而非 t-SNE 或 UMAP 等非线性降维算法，因为 PCA 几乎可以实时放置节点，不会产生延迟，而且更适合保留全局语义距离。我们仅计算主题节点的 PCA 坐标，而不计算内容节点的坐标。如果将内容节点纳入全局 PCA 投影，它们可能会因语义属性不同而偏离其父主题，从而破坏用户陈述的逻辑分组。因此，内容节点会以独特的径向排列方式放置在其父主题周围，以确保它们在视觉上仍锚定于其来源。

第二阶段 动态优化采用轻量级的基于物理的模拟来微调布局。如图 7C 所示，该过程会计算所有主题节点与内容节点两两之间的余弦相似度；如果相似度超过 $\tau = 0.3$这一阈值，就会在语义相关的内容节点与主题节点之间施加微弱的吸引力。我们通过实验测试确定了这一阈值。这种移动为跨主题连接提供了视觉线索，满足了用户识别不同子要点之间关系的需求。

5 评估

我们开展了一项有 12 位参与者参与的受控研究，以了解 Orality 的可供性并评估其在支持思维澄清方面的功能，同时将其与基于 ChatGPT 的基线条件进行比较。通过此次评估，我们旨在探究：

- 研究问题1：Orality 的表现与基线相比如何?
- 研究问题2：Orality 的各项功能是如何被使用和感知的?
- 研究问题3：使用 Orality 时出现了哪些工作流程和策略?
- 研究问题4：与基线相比，Orality 如何支持思维澄清?

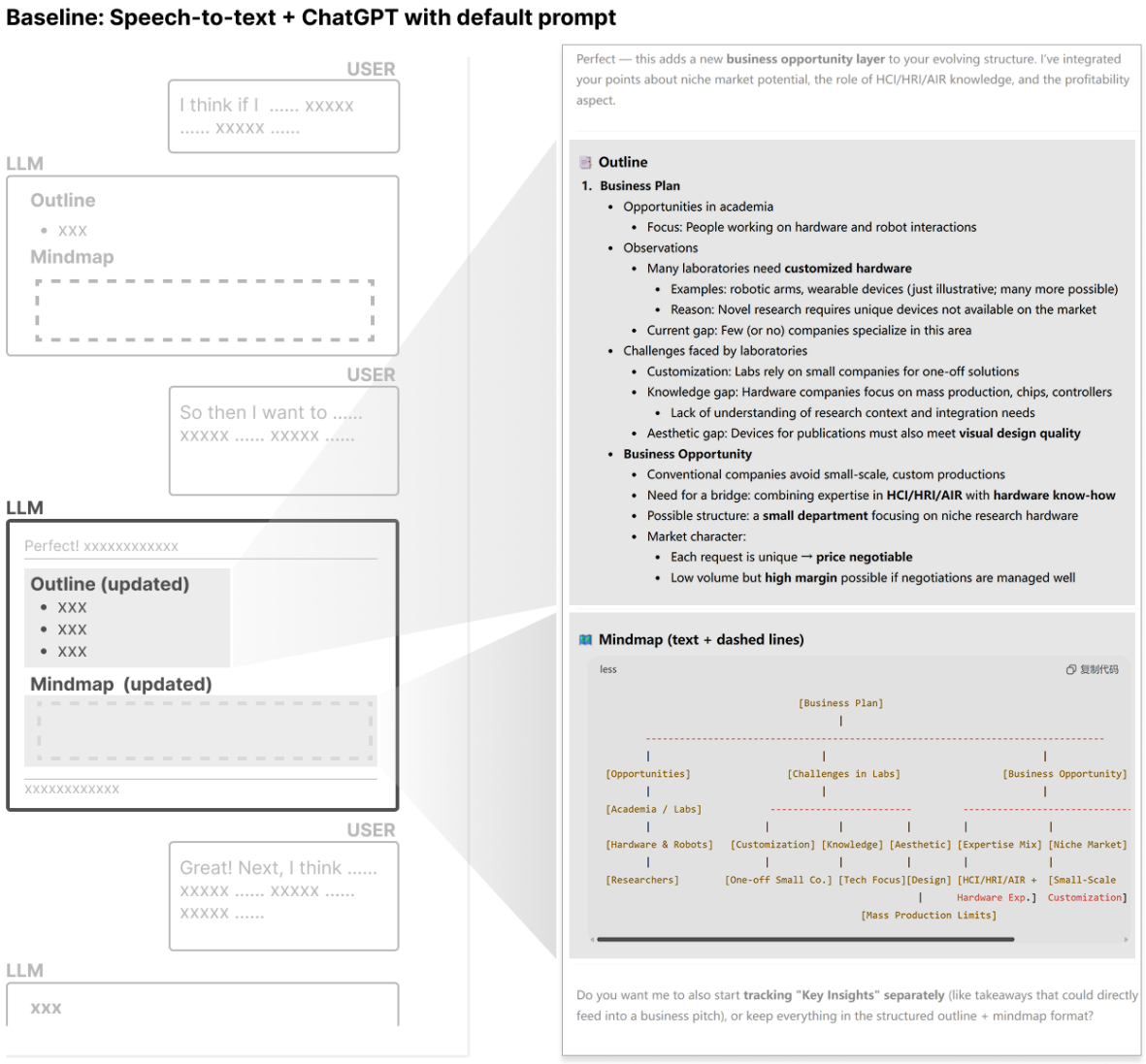

图 8：基线界面的示意图，展示了一段示例对话的屏幕截图。为支持思维组织，我们的默认提示词要求 ChatGPT 在每轮对话中生成一份大纲和一份基于文本的思维导图。

5.1 基线条件

基线条件是使用听写输入的 ChatGPT 界面，如图 8 所示，其设置与形成性研究相同。它提供了最先进的人工智能，并充当基于对话的思考伙伴，让用户能够将想法表达出来并加以整理。由于参与者向人工智能提供提示的能力各不相同，为确保在人工智能支持方面公平比较基线条件与 Orality，我们为 ChatGPT 编写了一个默认提示词，并在每次对话开始时注入该提示词。该提示词要求 ChatGPT 提供一份大纲，提炼参与者口头内容中的关键信息，并在适当时展示文本形式的概念图。与 Orality 类似，基线也有助于提炼和整理口头内容，同时通过生成的文本思维导图提供基本的视觉结构。参与者可以通过提交口述内容和后续提示词（例如“帮我梳理一下之前的内容”），参与者可以反复输入、组织和完善自己的想法。参与者还可以使用键盘和鼠标回看并编辑先前的条目。基线提示词和一段对话示例分别见附录 E.1 和附录 C.1。

5.2 参与者

我们通过滚雪球抽样，从一所当地大学招募了 12 名参与者（N=12）。据参与者自述，其中 6 名为女性，6 名为男性。所有参与者都能流利使用英语，且在语音转文本技术和生成式人工智能方面的使用经验各不相同。所有参与者均获得 20 美元作为参与补偿。参与者在语音转文本方面的经验各不相同

以及在生成式人工智能方面的经验也各不相同。参与者的详细信息见表1。我们还在附录 C.3 中报告了每位参与者提出的主题。

5.3 任务设计

参与者分别使用 Orality 和基线完成“思维澄清”任务。该任务旨在引导参与者连续言语化自己的想法，借助所提供的工具将想法外化并逐步加以完善，重点在于逐渐澄清思维过程，而非立即生成解决方案。相关指令见补充材料。

任务主题由参与者提出，以确保两种条件能够进行公平比较，即使参与者在两个主题上的熟悉程度和初始认知起点保持一致。每位参与者都被要求生成两个不同的主题，可以自行构思，也可以参考所提供的默认主题。所提出的主题应涉及一种让他们感到思绪混乱、需要通过外部表征加以澄清和发展的情境。补充材料中准备了五个主题示例，作为默认参考。参与者还被要求说明主题选择的理由，解释与澄清这些主题相关的挑战。经实验人员批准后，这些主题即确定用于后续任务。

5.4 流程

本研究为一次时长90分钟的会话，根据参与者的偏好，可在可用性实验室线下进行，也可远程进行，因为所有工具和流程均可通过笔记本电脑上的Chrome浏览器完整访问。该流程包括三个阶段：研究前准备、主要研究和研究后访谈。在线下环境中，会话在一间安静的房间内进行，仅有一名实验人员在场。

研究前准备。 一开始，参与者需要填写一份关于其人口统计数据的问卷，收集的信息包括年龄、性别、职业，以及以往对语音输入和大语言模型的使用情况。随后，他们被要求提出两个主题。当实验人员确认所报告的主题后，他们即可进入主要研究。

主要研究。 随后，参与者使用提出的两个主题分别在两种条件下完成了两项“思路澄清”任务。主题由实验人员随机分配。界面条件的顺序在参与者之间进行了平衡。在任务开始前，每位参与者都可以进行练习环节，以熟悉 Orality 和基线各自的功能。我们鼓励参与者逐步完善并梳理自己的想法，优先考虑深度和连贯性，而非尽快得出结论，但这并非强制要求。我们指示参与者主要将系统视为组织辅助工具，自行生成原创内容，并避免要求系统生成大量内容。此外，在两项任务中，我们明确不鼓励使用“帮我制定计划”等生成式开放式指令。

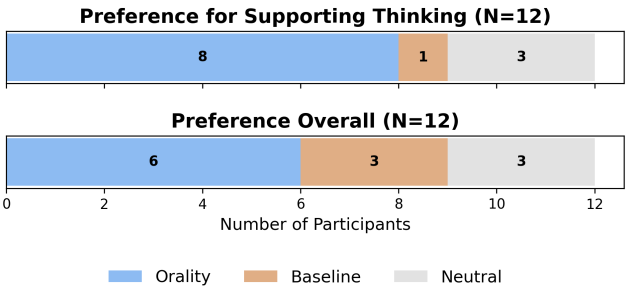

图9：参与者分别就支持思维过程和整体使用对Orality与基线的偏好。

建议参与者在20分钟内完成任务，但这不是硬性时限。时间到后，实验人员会提醒参与者；他们仍可继续完成任务，直到觉得自己的想法足够清晰。每项任务结束后，参与者需通过多个7点李克特量表对工具的有用性、对任务的满意度以及各项功能的质量进行评分，并填写NASA-TLX问卷，以评估任务期间的工作负荷。两次会话之间安排了短暂休息，以减轻疲劳效应。

研究后访谈。 参与者就其用户体验以及对两种条件的偏好接受了访谈。具体而言，访谈方案涵盖了对每种方法所感知的优势与缺点、起草过程中采用的策略、对内容的控制感、对Orality功能的看法与使用，以及其可用性、潜在应用场景和改进建议。详细的访谈方案见补充材料。访谈持续约25分钟。

5.5 数据收集与Analysis

实验期间，我们收集了多种形式的数据，以捕捉参与者的行为和经历。所有任务环节均通过全屏视频和音频录制。我们使用任务后问卷收集主观评分，问卷包含关于思维澄清满意度的7点李克特量表，以及NASA-TLX工作负荷量表。此外，任务结束后，我们进行了半结构化访谈，以收集参与者的反思。

对于定性数据（出声思维协议、转录文本和访谈回答），我们采用主题分析，以识别反复出现的模式，并提取不同条件下的关键见解。两位研究人员分别对三位参与者的访谈记录进行编码，并讨论各自的编码和初步主题，以达成一致。一位研究人员随后继续对其余数据进行编码。最终主题经研究人员反复讨论后确定。

对于定量数据（交互日志、视频分析和量表评分），我们计算了描述性统计，并对两种条件进行了统计比较。

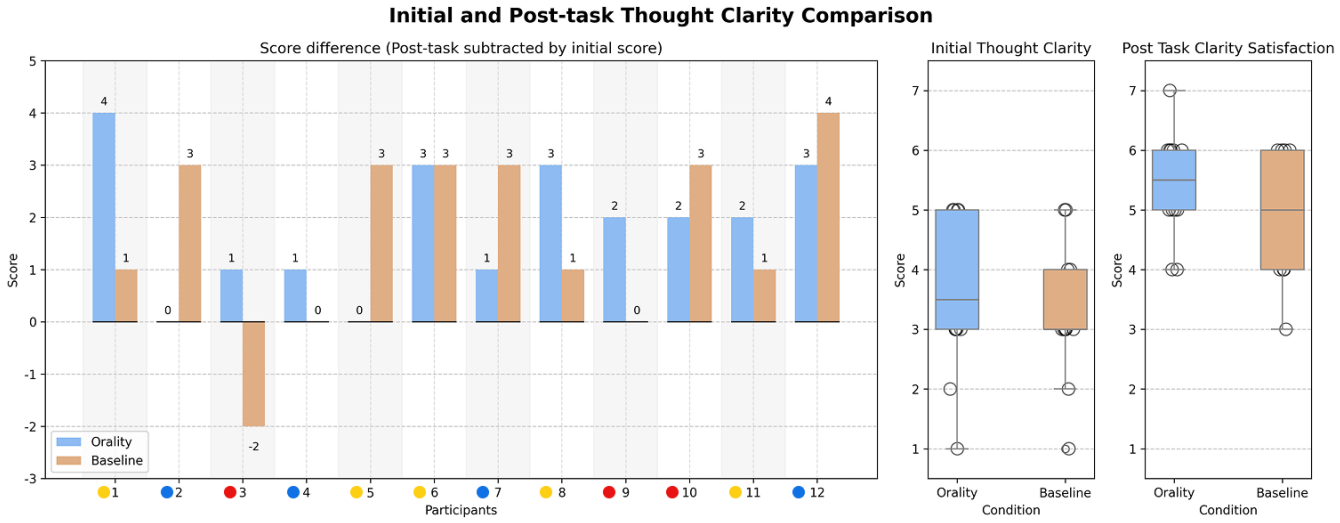

图10：任务开始前和结束后思路清晰度的评分。左：按参与者和任务分别展示任务开始前与结束后评分的清晰度差异。根据参与者在第6.3部分中识别出的工作流程，以不同颜色标示所属分组（黄色："指挥大师"；蓝色：亲力亲为的绘图者；红色：懒于操作的说话者）。右：所有参与者在任务前后给出的实际清晰度评分汇总箱线图。

6 结果

6.1 研究问题1：Orality与基线相比表现如何?

我们尝试通过两个李克特量表问题，衡量思维澄清任务中两种条件下的任务满意度。问题询问参与者在任务前后的思维清晰度。如图10所示，参与者分别对 Orality 的初始思维清晰度（M=3.58，SD=1.31）和基线条件（M=3.250，SD=1.14），以及 Orality 的任务后清晰度满意度（M=5.42，SD=0.90）和基线条件（M=4.917，SD=0.996）进行了评分。通过比较任务前后的清晰度差异，我们没有发现两种界面条件之间存在统计学上的显著差异，不过 Orality 条件呈现出清晰度较高的趋势。我们还观察到，采用不同工作流程的参与者之间存在一些有趣的差异，将在第 6.3 节中加以说明。

研究要求参与者说明他们对使用这些工具的偏好。如图 9 所示，当被问及哪种工具能更好地支持他们的思维过程时，12 名参与者中有 8 名选择了 Orality，1 名选择了基线。有趣的是，当被问及总体偏好，即更喜欢哪一种工具时，6 名参与者更喜欢 Orality，而 3 名更喜欢基线。这一差异可能部分源于参与者在任务中所持有的不同思考目标，详见第 6.4.2 部分。有 3 名参与者在这两项指标上都表示没有偏好。一些参与者指出，基线更适合某些特定主题。例如，P3 在处理具体问题时更偏好基线，但在从全局视角思考一般性问题时更喜欢 Orality。P7 认为，Orality 提供了一种将一般性想法转化为具体想法的结构，而基线则更适合小规模的闲聊类任务。P12 认为基线更适合实现特定目标，而 Orality 更适合进行更广泛的讨论，例如头脑风暴。

在主观工作负荷（NASA任务负荷指数）方面，我们未发现Orality与基线之间存在统计学上的显著差异。定量测量未显示统计学上的显著差异并不令人意外，因为为了保持外部效度，实验任务是一个由参与者自行提出的开放式思考任务。不同任务之间，任务选择和参与者的思考策略差异很大。我们将在第6.4部分通过定性分析继续比较Orality与基线，从而进行更有启发性的比较。

6.2 研究问题2：Orality的各项功能是如何被使用和感知的?

6.2.1 质量评估与有用性。 我们计算了各质量量表的描述性统计（图11）作为参考：口语内容的信息抽取（M = 5.08，SD = 1.16）、语义画布清晰度（M = 5.00，SD = 0.95）以及AI建议（M = 5.08，SD = 1.62）。可以看出，该系统的主要输出整体表现尚可，但个案之间存在差异。调查还评估了Orality各项功能的有用性（图12）。尽管大多数评分从"稍微有用"到"极其有用"均为正向评价，个体差异也很大。向我提问获得了最积极的评分，其次是口头结构化和语义空间。显示冲突似乎不太受认可。在访谈中，尽管大多数参与者认可其价值，但也报告了若干问题。

6.2.2 用于激发思考的人工智能建议。

向我提问通过就地提示激发新的、更深入的想法。几乎所有参与者都认为，AI 生成的启发式问题有助于在不同情境下支持他们的思考。许多

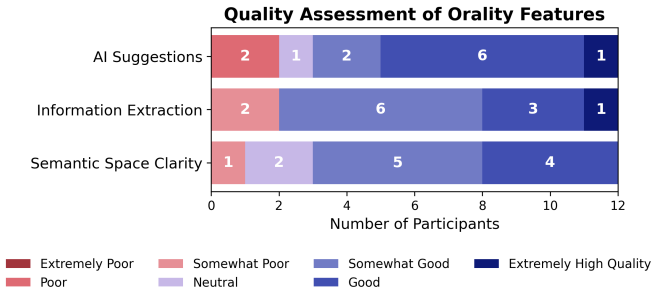

图11：参与者对 Orality 各项功能输出质量的评价。“信息抽取”和“语义空间清晰度”指系统从用户语音中识别主题并以可视化方式呈现的能力。“AI建议”指 AI生成的问题及检测到的冲突的相关性与质量。

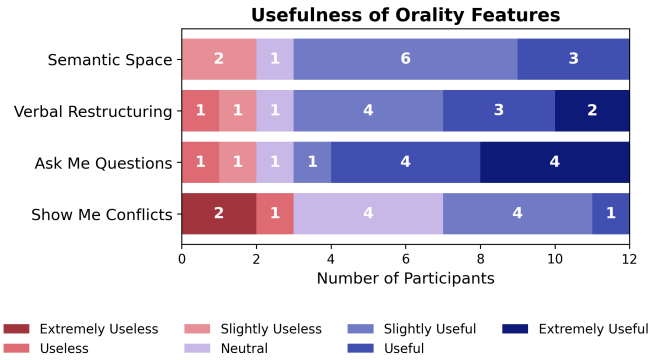

图1 2：参与者对 Orality 功能有用性的评估。

他们报告称，当不知道接下来该说什么时，这些建议尤其有用。例如，P7 惊讶地发现，一句随口的话触发了系统提出的一个意想不到的问题，并由此激发了一个新想法；而 P3 指出，在不确定该说什么时，它有助于打通思路。此外，参与者还发现它有助于阐述抽象主题：P2 解释说，系统提示他提供具体示例；P4 将其比作把想法拆解成“像树上的枝条一样”；P6 则用它来扩展子主题。P7 进一步表示，他很欣赏问题式建议而非陈述，因为它们开辟了新方向并鼓励更深入的反思，例如“我的目的是什么？”。在任务的初始阶段，参与者倾向于在阐述最初的想法后立即使用它来获得启发性反馈；例如，P9 解释说，早期的问题有助于澄清最初模糊的想法。在中间阶段，像 P12 这样的参与者更倾向于在已经对想法有一定熟悉度后再使用该功能，并发现此时提示词更为有效。

“显示冲突”功能虽有用，但仍需改进。P12 强调了该功能在识别两条可能路径之间的重叠或矛盾之处方面的价值，而 P3 则建议由系统自动检测冲突，而不是由用户手动触发。P8 对检测到冲突后缺乏解决冲突的支持感到困扰。P4 在系统未能检测出自己发现的冲突时，对系统产生了怀疑——这表明人们对“冲突”的含义存在模糊不清的理解。

6.2.3 口头结构化。参与者表示，在看到系统生成的初始聚类后，这项功能有助于整理思绪。它让用户能够自由地通过口头表达重新构建想法，从不同视角审视自己的思维过程，同时执行高级操作。对用户指令的分析揭示了四种截然不同的使用模式。

最常见的模式是进行具体的、高层次的组织。参与者利用这种方式，为无结构内容赋予清晰的主题结构。例如，P5 指示系统：“将我的笔记组织成三个部分：要求、框架信息和相关信息，”而 P7 也发出类似指令，自上而下地将内容重组为“背景、挑战和方法。”

参与者还使用合并或拆分指令来控制信息展示的粒度。他们并非要定义一种僵化的新结构，而是让人工智能改进现有结构。这包括整合信息，例如 P10 要求工具“合并我的一些想法，让图表更清晰，”，或 P3 请求工具“将‘相关信息’部分按不同任务类型划分。”正如 P1 所解释的，面对过多类别时，使用指令将它们合并能让他获得新的视角，并加深对自己想法的理解。当初始布局与他们的心智模型不符时，P10 会指示系统将一个父节点拆分为多个子节点。

第三种模式是使用系统综合信息以回答具体问题。在这些情况下，用户将系统视为能够解读自身数据的分析伙伴。例如，在口述完自己的职业经历后，P2 询问系统：“告诉我为什么我适合用户研究员职位，”这要求工具找到并分组相关节点，以提供可视化答案。

最后，参与者将口头结构化 用于高级分析和格式整理任务，拓展了该功能设计能力的边界。诸如 P12 要求“找出这些路径之间的……冲突点”，或 P8 指示“围绕一个主题总结四个项目”等指令，体现了参与者希望系统执行更高阶分析的意愿。

6.2.4 语义画布：思维澄清与发展。

帮助监测并发展思维。 参与者强调，从口语输入中提取主题有助于减少冗余和重复，而按颜色和位置对相关节点进行分组则创建了一个结构化画布，充当思维块。P1将其描述为“让思维清晰地呈现在画布上”，这帮助他在后续阶段回忆先前的内容。同样，P11指出，整齐排列节点使他能够判断想法的清晰度：分支越多表明清晰度越高，而分支较少则促使他利用后续问题进一步提炼想法。

空间信息的个性化含义。 参与者积极操控语义空间 以检视思维块之间的关系。他们利用空间信息来表示关系和优先级。例如，P12将画布分为左右区域以比较簇，而P9将相关节点放在一起，或将子主题围绕中心节点排列。P9还将重要节点置于中心、补充内容置于下方、较不重要的节点置于边缘，从而对节点进行优先级排序。

6.2.5 用于探索替代方案的思维演变。参与者使用这一功能回顾并恢复早期版本。12 位参与者中有 7 位使用该功能查看先前版本，这帮助他们找回遗忘的思路线索、比较不同版本的结构，或恢复已删除的内容。P5 解释说，重新查看早期布局让他得以回想起最初的版本并找出其中的变化，而 P11 和 P12 则强调，该功能在新生成的布局不尽如人意或存在错误时，可作为撤销功能使用。参与者还表示，回顾和撤销更改的能力让他们敢于尝试新的结构，不必担心丢失有价值的内容或打乱自己的逻辑。例如，P2 在完善访谈回答时，于很短的时间内多次使用了这一功能。通过恢复到先前的状态，该参与者得以针对自己的工作经历尝试不同的叙事结构，再选定最终版本。P8 描述了如何利用这一工作流程尝试新的方向：他们会口述一个新主题，判断它是否“特别好”，如果不是，就返回先前版本，思考如何改进。除了反复试错之外，对上下文的回顾也支持对逻辑流程进行元认知监控。P9 解释说，回顾历史记录有助于“理清整个逻辑流程”，让他们看到自己的逻辑截至当时是如何组织起来的。

6.2.6 导出完整答案。最后，只有少数参与者使用并认可导出 功能，用它将思维澄清的过程总结成文本报告；他们认为这是互动带来的有用成果。P1 指出，这“更像是在记录我之前的想法”，而 P10 则将其描述为“

以后可以用作备忘录的内容。”其他人则强调其在多次迭代后澄清零散想法方面的价值，P8表示该报告“曾帮助我一次性通览所有内容，并使想法更加清晰。”

6.3 研究问题3：使用 Orality 时出现了哪些工作流程和策略?

对观察数据（系统日志、屏幕录像）的分析揭示了与 Orality 交互的三种 distinct 模式。本部分通过补充参与者本人描述和解释中的引文，对这些模式进行说明。我们根据参与者对输入模态的使用方式、交互模式，以及他们用于人机思维组织的底层心智模型，对参与者的工作流程进行分类。由于参与者的任务策略本质上受到其个人风格和选择的影响，我们将参与者分为三组进行综合和报告：“指挥大师”、亲力亲为的绘图者和懒于操作的说话者。附录 D 中列出了每位参与者工作流程的详细描述。

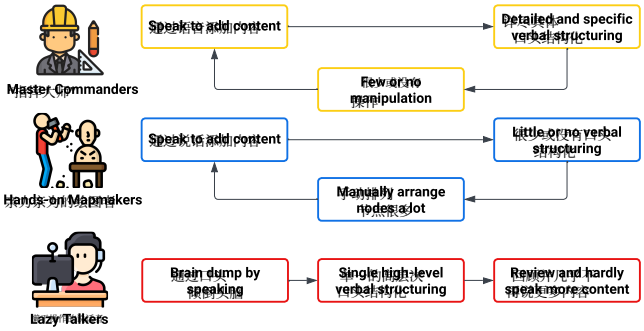

图 13：使用 Orality 时观察到的三种交互策略和任务工作流。

6.3.1 “指挥大师”（*P2*、*P5*、*P7*、*P8*、*P12*）。第一组也是人数最多的一组，几乎完全通过内容听写和一系列具体且通常较为复杂的语音指令进行操作（图 13 顶部）。这一组的典型行为是在整个会话中使用大量的口头结构化指令，同时几乎不进行或仅进行极少的手动操作。他们通过一连串指令来实现目标，并大量使用思维引导式问题。他们并不只是组织内容，而是在积极地与 AI画布 进行对话，以构建、综合，甚至分析自己的想法。

例如，P5通过口述一系列研究框架和任务类型来开启会话；当初次重组并不完全准确时，P5并未手动整理这些节点，而是使用结构命令来定义地图：“我想把我的笔记分成三个部分。第一部分是需求……第三部分是所有框架。”在会话后期，P5又使用了另外两条指令：“相关信息过于密集，请按不同任务类型在每个节点中进行划分”以及“合并重复的信息。”这表明了一种工作流程：用户将AI画布视为可通过迭代口头指令引导的组织助手。同样，P2使用一系列口头结构化操作，将其学术与职业经历整理为清晰的类别（例如，“研究助理经历”、“专业工作经历”）。P7的过程具有明显的层级性：首先，他们下达指令，将所有内容自上而下地广泛重组为一个分析框架（例如，“背景”、“挑战”、“方法”）。随后，P7“放大查看”，选择一个新的部分，并再次发出更细粒度的指令，以重组其内部布局。

这些用户选择这一工作流程，是因为他们认为，在结构化复杂想法方面，语音是比直接操作更强大、更准确、更高效的媒介。他们几乎完全通过口头结构化 来组织内容，因为这使他们能够执行那些若靠拖拽和编辑节点就会变得复杂乃至无法完成的结构性与分析性任务。这一点在P12的日志中十分明显：P12不仅在空间上组织了主题，还要求系统执行逻辑分析。他们指示系统在其两条职业路径之间“找出中断部分……重叠部分……以及冲突部分”。后来在访谈中，P12解释说，这是因为其分析目标有此需要，并指出：“我让他列出两条路径之间的重叠部分……我想清楚地看到哪些部分发生了重叠。”

6.3.2 亲力亲为的绘图者（*P1*、*P6*、*P9*、*P11*）。第二个分组展示了一种工作流程，它依赖于语音与直接操作的紧密结合（图13中部）。他们的标志性行为包括以简短、迭代式的突发方式使用语音，随后进入以手动重组为主的阶段，以创造空间意义。他们以极低的频率使用口头结构化，更倾向于手动结构化自己的想法。

例如，P9 主要将该工具视为一个可视化项目规划画布。在全面介绍项目组成部分后，P9 进行了细致的手动排列，连续执行了七次主题拖动操作，将“项目合规”“协作”和“即时项目需求”等簇整理到空间的三个角落。P9 这样描述这一排列过程：“我刚开始排列内容时，会先优先处理最重要的关键词，然后从上到下进行排列……我会把主要分支的小气泡移到两侧，以便更轻松地阅读。”类似地，每次添加少量内容后，P1 都会花费大量时间手动排列节点，利用空间布局建立联系并梳理思路。P1 解释了自己逐步进行手动分类的方法：“我先把它们分开，然后逐个阅读，看看哪些是新生成的，哪些是后来生成的。读完之后，我决定根据自己的想法进一步对气泡进行分类。”

选择采用这一工作流程，源于他们相信思考是一种具体、视觉化且具有空间性的活动。这些用户并不完全信任人工智能的逻辑能够呈现自己的想法，或者觉得亲手操作更有意思，因此他们依靠直接操作，通过亲自摆放来塑造想法，并借助布局赋予其意义。正如 P9 所解释的：“我认为布局有助于理解。因为一开始可能会有一些问题……我可能会先把它们放在最初输入的内容旁边，进行比较……我觉得安排这些内容的过程很有意思。”他们觉得自动布局令人困惑或受到限制：“我让它帮我重新整理并删除重复内容，结果发现只剩下三点，……这让我觉得有点不踏实。”（P9）以及“我希望它能这样做，因为还有很多重叠的内容，……但似乎没有起作用。”（P11）对这些用户而言，雕琢画布的过程与理解的过程密不可分。

6.3.3 懒于操作的说话者（*P3*、*P4*、*P10*）。最后一组采用的工作流程以线性两步过程为特征：先把所有内容说出来，再发出少量指令（图 13 底部）。他们的行为特征是在会话开始时大量使用语音，类似于“头脑倾倒”，几乎不进行手动操作。

P10 清楚地体现了这一模式：他们在会话开始时连续输入了三段长内容，涵盖职业困境、博士学位与工作的优点和缺点等，之后才发出一条全局指令：“把我的一些想法组合起来，让图表更清晰。”P4 也体现了类似的行为。在对当前生活的各个方面和职业不确定性进行了一番长篇“头脑倾倒”之后，他们要求人工智能调整对话的结构，专门聚焦于“毕业后不确定性”。此外，P3 开始时口头概述了一个明确的问题（一篇被拒的学术论文）及其背景。随后，他们继续通过口述详细说明具体的审稿人反馈，并集思广益地探讨可能的解决方案。P2 也描述了这一漫长的初始“倾倒”阶段；他们认为，仅仅将想法说出来就很有效：“我只是想先描述一下我自己……我先把脑海中浮现的一切都说了出来，然后觉得自己部分达成了目标。”

访谈显示，这种行为源于清晰的成本效益分析。这些用户觉得手动整理的过程“代价高昂”，并认为视觉布局与他们的最终目标无关。P10 明确表示，手动布局需要付出很大努力，因此他们作出了这样的选择：“我觉得布局方式用起来相当费劲……我觉得整理起来太麻烦了，所以干脆没做。不过总体来说，我能理解它试图表达的层级结构。”P4 解释说，由于视觉布局与自己的目标无关，他们没有进行操作，并表示：“我把它们都看了一遍，但没有整理，因为我觉得它们都由一个节点连接着。我认为整理与否并不会有什么区别。”

6.4 研究问题4：与基线相比，Orality 如何支持思维澄清?

本部分报告了我们的研究发现：根据对访谈数据进行的主题分析，我们探讨了所测试的工具如何支持参与者通过思维过程实现各自的目标。

6.4.1 支持发散性思维：引导持续拓展与提供额外信息。我们的分析表明，Orality 和基线都支持发散性思维，但方式各不相同。研究发现，Orality 通过引导用户多加思考，为他们的思维提供支架；而基线则因提供外部知识并让思绪自由流动而受到认可。P5 描述了 Orality 如何通过提取主题并提供 AI建议来搭建“支架”：“……它会引导我思考分类，即使分类可能不正确，我也可以通过提供指令来反复调整。然后它还会给出 *AI*建议并检测冲突，这些都会引导我思考额外的想法，我觉得这很有帮助。”P9 也表示，Orality 提出的问题“帮助我跳出脑海中已有的内容，拓展思路。”另一方面，基线提供的外部信息也受到了认可。P2 在围绕国际学生如何省钱这一主题进行头脑风暴时，喜欢基线补充的其他省钱方式。P5 使用基线制定旅行计划时，基线为其补充了住宿和餐饮信息。P5 评论道：“（基线）提供了一些额外信息，帮助我输入更多内容来完成计划。”P1 阐述了自己使用这两个系统时不同的心智模型：“（*Orality*）会让你解决每个小问题，从而逐步解决大问题，它是一个问题解决工具。使用基线更像是在进行头脑风暴，你不停地思考，系统则会为你总结。”P6 也提到了这种自由思考的状态：“使用（*Orality*）时，我可以跟随它提供的主题进行表达，而（基线）并没有太多组织内容，可能会更自由一些。”

6.4.2 支持意义建构：鼓励主动意义建构，而非直接提供答案。首先，两个系统都通过说话这一行为来支持意义建构。正如 P9 所说：“说话时，我的大脑已经把内容过了一遍……说出来帮助我理清了要做的事，因为我必须把它描述出来。”随后，观察到的最显著差异是，Orality 引导参与者得出自己的结论，而基线则通过提供答案来提供帮助。P7 对此阐述得很清楚：“（基线）倾向于

提炼精髓，而（*Orality*）则构建完整的框架。我说的'精髓'是指两个要点，也就是最后我真正需要的内容。" P7 还表示，借助 Orality 构建的框架引导他们"从模糊走向具体"。基于这些观察，P7 得出结论：对于较小或更主观的想法，基线更有帮助，因为它能直接抓住核心；而 Orality 则适用于更复杂的内容。类似地，P6 说："（*Orality*）适合用于分析，而（基线）会整理想法并给出它的意见。"也许这正是 P9 自己注意到的一种行为模式：他们大多通过指出"哪里有问题"来"纠正"基线，而使用 Orality 时，则会依据 AI生成的问题"扩展描述"。

此外，一些参与者表示，与基线相比，他们使用 Orality 进行了更多迭代。P4 用这两种工具对比了它们的差异：Orality"促使我一点一点地多说一些"，因此结果更加"完整"；而基线"完全取决于我能从脑子里挤出多少内容"。P3 也认同这一点，并指出，相比之下，他们倾向于在基线中"一次说得更多"，而该系统则支持"迭代过程"。

尽管基线也提供思维导图来总结并组织用户的内容，但缺乏交互性和视觉连贯性似乎是其主要缺点。基于聊天的思维导图虽然基础，却提供了层级结构，一些参与者认为这很有帮助。P6 将其描述为"半成品"，但也承认它有助于组织内容；P10 则欣赏它能够多层次地扩展内容。P2 对于无法通过提示词修正基线概念图中的某些冲突感到沮丧，尽管他们认为它有助于分类和补全信息。P9 表示，与基线中的概念图相比，他们对 Orality 的不准确之处更能容忍，因为概念图不容易编辑。P1 表示更喜欢 Orality 中的语义画布，因为它"引导我深入思考"，而基线中的思维导图"更像是一份总结"，并没有"为我提供思考方向"。

然而，参与者对更主动的意义建构有着不同看法，具体取决于他们的任务目标和偏好（所有任务主题列表见附录 C.3）。当参与者希望人工智能直接提供解决方案或改善结果时，他们更倾向于基线。例如，P3正在寻求问题的解决方案，但觉得 Orality 没有用，因为"如果我是在做内容分析，那么将内容分类到不同主题中会有帮助。但今天我是在描述自己的困难，所以这些主题没什么意义。"P12赞赏基线能够直接"很好地完善了我的想法"。P2也更喜欢使用基线，因为它"降低了心理负荷"，并表示 Orality 让他们"思考得更多"，因为"我需要判断这种分类是否合理，并思考如何优化它"。P4也不喜欢 Orality 无法帮助他们"收敛"，觉得它更像一个"分析器"，而他们想要的是想法。

6.4.3 支持深入思考：持续向下探索，而非在同一层级间跳跃。据报道，Orality 有助于拆解复杂问题，促进思路"纵向"深入。参与者将其与基线进行了对比，指出标准的大语言模型交互往往只会生成"横向"摘要，罗列并列的概念，却缺乏深度。P1 描述说，基线会让他们的思考停留在同一层面，例如比较城市的总体优缺点；而 Orality 则能引导他们将"高层次的目标，例如"如何成为一名软件架构师"，分解为具体可行的步骤，比如从"后端开发"到"服务器开发"，再到"特定编程语言"，最终细化为具体的学习资源。

参与者表示，Orality 支持了更流畅的思考过程：视觉布局帮助用户追踪已经外化的内容并规划下一步，从而让思考更加聚焦。当思考被打断时，参与者可以通过扩展内容不足的簇（P1、P10）或继续探讨感兴趣的子主题（P6）来接续思路。相比之下，基线有时会打断思考；当对话变得冗长且文字密集时，参与者会感到困惑（P1）。这要求用户付出额外努力，从冗长的对话记录中梳理并提取要点（P4、P7、P8），还要辨识相较于先前交流有哪些新进展（P4）。

6.4.4 支持思想发展：有针对性的思考与伙伴式聊天。 据报告，Orality 能有效支持专注且有针对性的思想发展，尤其得益于其主题提取和节点选择功能。P1 描述称，其口述内容的初始聚类"让我始终聚焦于这两个主题"，并感到随后的思想拓展是基于它们展开的。P6 对此表示认同，称使用 Orality 时他们"围绕一个主题发言"，而使用基线时则"更为发散"。

能够对选定的节点执行操作，帮助参与者以更强的控制来发展目标领域。P8 强调了这一点的重要性，这样他们就可以"针对某个目标进行思考或提问，.....这有助于把一件事分成几个部分来思考，或将其作为上下文来请求系统建议"，同时将基线对比为"在我这边和系统那边都很混乱"。P6 充分利用了这一功能，强制规定了一系列人工智能操作的顺序。他们先对每个主题使用"向我提问"功能，然后再用"显示冲突"检查潜在冲突。P10 认为 Orality 正是因此能更好地支持思考："它基于特定节点提问，这使得层次清晰。但如果（基线）向我提问，我就不知道它指的是什么"。

另一方面，基线的基于聊天的界面也带来了一定的参与效果。一些参与者很看重那种能给予肯定的"如同伙伴般"的感觉。正如参与者 P1 所说，当他们想到的下一步与基线提出的问题相吻合时，他们会更有信心："因为这种来回讨论会让你觉得自己是在和伙伴交流；它会回应我，让我对下一步更有把握。"

6.5 用户建议的改进和潜在使用场景

解决节点过多导致的信息过载。 参与者表示，节点数量过多会造成视觉杂乱和认知过载。他们提出的解决方案侧重于界面设计和自动化。界面方面的建议包括使用预定义布局模板（如树形结构）来系统地组织节点（P4）。为降低信息密度，P9建议自动提取并突出关键词，以便更快地浏览。另一方面，参与者也建议通过自动化实现实时合并和分类。例如，P1希望将内容自动整理到用户定义的组中，而

P3建议在输入过程中对相似内容进行集群，以防止空间过度拥挤。

提供更多节点交互。 参与者还希望节点交互更加灵活且具有语义意义。他们希望能够创建或修改不同节点之间的关系，例如通过视觉连接将设计示例与理论框架关联起来（P5）。另一项重要建议是语义缩放，使用户能够将特定节点（例如“城市 X 的优势”和“城市 Y 的优势”）折叠到更宽泛的父类别中，例如“城市的优势”（P6）。最后，用户希望除了当前基于时间的颜色编码外，还能从多个视角对节点进行分组。例如，P7建议将“方法”或“挑战”中的内容重新归入优势或局限等新类别。

建议的潜在用户场景。 根据参与者反馈，Orality非常适合三种主要场景。第一种是构思与反思性决策，例如头脑风暴（P12）或模拟人生选择（P4）。第二种是规划与全面的信息组织，例如创建旅行清单（P9）或回答复杂的比较性问题（P6）。第三种是为学术与定性研究组织信息，尤其是管理非结构化数据的多层编码（P3）。总之，该工具最有益于需要探索性思维、多层次分析，以及对最初模糊的想法进行结构化输出的任务。

7 讨论

我们的研究表明，在语音驱动的思维澄清方面，Orality相较于对话式大语言模型界面具有独特的认知与交互优势。Orality并非仅仅复制可视化画布 [30, 41, 54], 的已知益处，而是引入了一种语音优先、具备空间感知能力的意义建构循环，将流畅的言语表达与结构化的视觉呈现独特地结合起来。下文中，我们从认知卸载、元认知支架和交互模态权衡的视角解读研究发现，并推导出超越以往人工智能增强型思维工具研究的设计启示。

7.1 语音：结构化思维的主要模态

以往的系统将语音用于口述或命令，而 Orality 则将语音视为持续、结构化构思的原始素材。这超越了 Bolt 奠基性的“放到那里”[4] 范式，不再将语音仅仅视为控制机制；也超越了 Rambler [25] 或 MeetMap [6] 等专注于生成最终成品的工具。我们观察到的用户工作流程——尤其是“指挥大师”——表明，参与者使用语音不仅是为了添加内容，也会通过元指令重新构想内容（例如，“将这些合并”、“将这些内容整理成这三个部分”.....）。这说明，当语音与响应灵敏的视觉呈现相结合时，便可成为迭代式自我对话和概念重组的有力工具。我们观察到的三种不同工作流程表明，语音可以发挥多种作用：作为人工智能介导重组的指令工具，作为手动空间构建意义的补充，或作为门槛较低的倾倒式输入机制。这种多样性表明，语音可以成为一种灵活的模态，适应用户的认知风格和任务阶段。

7.2 人工智能的角色：从对话伙伴到嵌入式支架

人工智能增强型思维工具的一项关键设计挑战，是在提供有益的自动化与保留用户自主思考之间取得平衡。我们的工作置于不断演进的人机交互领域思维工具讨论背景下 [43]。近期研究显示，让人工智能提问 [46] 或根据用户自身的思考方式进行引导 [31]，也能带来类似的积极效果。Orality 的设计采用了一种独特的方法：将人工智能辅助直接嵌入用户图式化思维的空间表征中，而非将其置于独立的对话渠道中。用户高度评价的原位“向我提问”功能，充当了用户自身概念图中的认知探针。这将人工智能的角色从可能打断用户叙述的对话伙伴（这是基线的一个已知缺点），转变为拓展并深化用户思维的支架。

现有的人工智能辅助画布界面研究已显示，空间表征有助于发散性思维，并能增强监控与导航支持 [30, 41, 54]。在此基础上，我们的工作以语音优先工作流为特色，通过语音输入与可灵活调整的空间工作区实现实时联动。它支持在 Pirolli 和 Card 的意义建构循环中，频繁切换于搜寻（外化）和图式化（结构化）阶段 [29]。我们的研究发现，这种方法能够通过促进发散、深入的思考、主动意义建构和有针对性的思想发展来支持思考。

与近期倡导“生成性摩擦” [26] 或设计能够激励人们思考的人工智能 [33],的讨论一致，我们的研究表明，即使提取的主题不准确（第 6.4.2 节），Orality 仍能通过引导用户思考，支持更主动的意义建构过程；而基线则直接为用户提供答案。12 位参与者中有 8 位认为我们的系统更能支持思考，这表明此类方法颇有前景。

7.3 设计启示

用户研究的发现促使我们为未来通过语音和语义界面支持思维澄清的工具提出设计启示。

7.3.1 构建自适应语义表征。 虽然我们的语义画布已被证明有效，但参与者的反馈指出，表示方式还有更多样化、动态化的空间。一些用户觉得手动布局“成本高”（P10），另一些用户则希望空间关系更具表现力，例如跨类别连接节点（P5），或从多个视角创建不同的分组（P7）。这表明，具备基本信息抽取和图式化功能的简单自由形式画布，可能无法满足所有任务或用户的需求。未来的系统应支持多种表征视角（时间线、层级结构、网络、象限），并能根据任务情境切换或合并这些视角。用户应能够按照不同的语义布局（例如按主题、优先级或时间顺序）重新安排相同内容的空间布局，以支持多视角的意义建构。未来

的系统应探索其他形式及混合型语义空间，例如：

- 预定义模板：提供可选布局（如时间线、层级结构、象限图），为常见思维模式提供支架。
- 语义缩放：允许用户在不同抽象层级上折叠和展开节点组。
- 多视图：支持用户通过不同的结构视角（如按主题、按优先级、按时间顺序）对同一组想法进行可视化和交互，从而开展更深入的分析。

7.3.2 设计情境化 *AI* 建议。我们的研究表明，AI 生成的启发式问题（“向我提问”功能）对于激发新的、更深入的想法备受认可。在情境内生成的问题和提示词被放置为尚未充分发展的主题领域中的潜在新节点，有助于以可视化方式突出显示思维空白。这使 AI 建议从突兀且脱离情境的提示，转变为融入情境、理解上下文的元素；这些元素能够顺着用户已建立的空间逻辑自然延展，从而降低认知负荷，并支持更连贯的思维过程。“显示冲突”功能的使用率较低，这揭示了检测和展示复杂语义关系方面的设计挑战与机遇。未来研究可探索引导用户解决冲突的方案。此外，虽然我们的系统仅探索了空间情境，但参与者希望 AI 能更加主动。未来研究可纳入时间维度，为思考提供更具情境性的 AI 支持。

7.3.3 提升口头结构化的意图识别能力。口头结构化功能的强大之处及其频繁使用，尤其是“指挥大师”（P2、P5、P7、P8、P12）对该功能的频繁使用，凸显了用户对基于语言进行高层次重组的强烈需求。参与者使用指令，不仅进行基本分组，还执行高级分析任务，例如综合内容以回答特定问题（P2：“告诉我为什么我很适合”），或寻找概念之间的重叠部分（P12：“列出重叠的部分”）。然而，当系统的解读或执行结果不符合他们的意图时，用户也会遇到局限。这表明需要通过更强大、更可靠的方法来改进这一功能，或许可以同时采用界面和技术解决方案，让用户更轻松地执行复杂转换，例如将整个画布转向新的视角、合并较大的部分，或自动应用用户定义的模板结构。

7.3.4 增强元认知支持。元认知方面的要求，部分源于言语重构的结果可能难以预测，这是由用户输入的自由形式所致。通过为常见操作提供快捷方式或用户定义的可重复使用的提示词，或许可以改善这一点。我们的系统已展现出提供一定元认知支持的能力，可帮助澄清含义、把握主题思想发展的进程、追溯思维演变历程并生成不同视角。未来的系统若要进一步增强此类支持，这是一个重要方向。例如，我们可以设计一种功能更强大的思维演变工具，允许用户通过用户定义的维度并在多个抽象层级上检查思维。系统还可以允许用户创建命名的检查点、分支出不同的思维轨迹，或筛选历史记录，仅显示与特定主题相关的变更。这将更好地支持元认知过程，帮助追溯想法的起源、理解不同决策如何导致不同结果，并在不丢失当前思路的情况下重新查看已放弃的路径。多位参与者使用了带有所选主题的导出功能，但使用不多，因为冗长的输出阅读成本较高。未来的系统或许可以改进这一功能，让输出更便于查看，并帮助用户发现多个版本的导出内容之间的联系。

7.4 局限与未来工作

本研究存在若干应当承认的局限。

7.4.1 实验室环境可能限制真实感与任务真实性。 首先，作为一项实验室研究，实验环境缺乏生态效度，参与者自行提出的思维澄清任务可能略显肤浅。未来研究应探索如何在更真实的任务与现实情境中，借助此类工具支持思维。

7.4.2 内容创建的差异性限制了比较。 虽然允许参与者自行生成任务主题对于鼓励批判性思维过程是必要的，但这也引入了任务难度与目标取向方面的差异。一些主题旨在得出具体解决方案，例如制定旅行计划，而另一些则强调对问题理解的增量式进展，例如个人职业路径。这种差异难以控制，并可能影响参与者对工具在澄清其思维方面有效性的感知。

7.4.3 系统输出不稳定。 我们的原型有时未能反映参与者预期的组织方式，原因包括内容上下文不足、语音指令不清晰以及大语言模型的不稳定性。研究中观察到错误分类、节点被意外合并或删除，以及冲突检测不完整等情况。参与者经常报告，系统未能捕捉其预期组织方式的细微差别，尤其是在“向我展示冲突”功能中，该功能有时未能识别出对参与者而言显而易见的冲突。未来工作应改进人工智能支持的技术性能。

8 结论

本文介绍了 Orality，这是一种人工智能增强的有机界面，旨在支持使用语音和语义画布来澄清思维。针对我们形成性研究中发现的挑战，Orality 提出了一个四层框架，用于：(1) 将线性语音转换为空间化的语义表征；(2) 提供口头结构化功能，以灵活地重组内容；(3) 通过按需、就地提供的 AI建议，激发更深入的反思与阐述；(4) 将思维过程的迭代演变可视化。用户研究表明，参与者能够成功使用 Orality 来澄清思维；与提示词增强的大语言模型相比，他们更倾向于使用 Orality 来支持思维过程。这展示了借助空间AI界面增强语音驱动的思维澄清的潜力。我们展望这样一个未来：语音优先、基于画布的界面将言语表达的流畅性与视觉语义的结构力量融为一体，最终

增强我们思考、创造并理解最复杂想法的能力。

致谢

本研究获得华为技术有限公司捐赠（项目编号：9220140）以及 Google 有限责任公司的 Google 教师研究奖（项目编号：9229068）资助。我们感谢 Alvaro Cassinelli 和 Advait Sarkar 富有启发性的讨论。衷心感谢匿名评审提出的建设性且富有见地的意见。为确保透明度，我们还想说明共同第一作者的分工：Wengxi Li 确定了理论框架，并完成了系统的设计与实现。Jingze Tian 主导了形成性研究和评估研究的设计与实施。两人均参与了数据分析和论文撰写。

我们承认在研究人员仔细审阅、确保原意不变的前提下，使用了 ChatGPT 对全文措辞进行润色并使其更加简洁。作者对内容的正确性和真实性负全部责任。

参考文献

[1] Eric A. Horvitz。1999。混合主动交互。*IEEE Intelligent Systems and their Applications* 14, 5 (1999), 14–23。doi:10.1109/5254. 796083[2] Barry Arons。1997。SpeechSkimmer：一种用于交互式浏览录制语音的系统。*ACM Trans. Comput.-Hum. Interact.* 4, 1 (1997年3月), 3–38。doi:10.1145/ 244754.244758[3] Sumit Asthana、Sagi Hilleli、Pengcheng He 和 Aaron Halfaker。2025年。摘要、要点和行动事项：基于大语言模型的会议回顾系统的设计、实现与评估。*Proc. ACM Hum.-Comput. Interact.* 9, 2, 文章CSCW 176 2025 29 10 1145 3711074 (2025年5月), 页数。doi: . / Richard [4] A. Bolt。“放到那里”：图形界面中的语音与手势。载于第七届计算机图形学与交互技术年度会议论文集(美国华盛顿州西雅图) *(SIGGRAPH '80)*。美国计算机协会，纽约州纽约，美国， 262 270 10 1145 800250 807503 – 。doi: . / 。[5] Senthil Chandrasegaran、Chris Bryan、Hidekazu Shidara、Tung-Yen Chuang 和 Kwan-Liu Ma。2019年。TalkTraces：会议中口头内容的实时捕捉与可视化。载于*2019*年*CHI*计算系统中的人为因素会议论文集(英国苏格兰格拉斯哥) *(CHI '19)*。美国计算机协会，纽约州纽约，美国， 1 14 10 1145 3290605 3300807 – 。doi: . / 。[6] Xinyue Chen、Nathan Yap、Xinyi Lu、Aylin Gunal 和 Xu Wang。MeetMap：在在线会议中使用大语言模型进行实时协作式对话映射。*Proc. ACM Hum.-Comput. Interact.* 9, 2, 文章 CSCW132 (2025年5月), 35页。doi:10. 1145/3711030[7] Michelene TH Chi、Nicholas De Leeuw、Mei-Hung Chiu 和 Christian LaVancher。1994年。引出自我解释有助于提升理解。*Cognitive science* 18, 3 (1994), 439–477。[8] John Joon Young Chung 和 Max Kreminski。2024年。Patchview：借助生成式尘埃与磁铁可视化、由大语言模型驱动的世界构建。载于第*37*届*ACM*用户界面软件与技术年度研讨会论文集(美国宾夕法尼亚州匹兹堡) *(UIST '24)*。美国计算机协会，纽约州纽约，美国，文章 77, 19 10 1145 3654777 3676352 页。doi: . / 。Andy Clark [9] 和 David Chalmers。《延展心灵》。*Analysis* 58, 1 (1998), 7 19 3328150 – 。http://www.jstor.org/stable/ Taufiq [10] Daryanto、Xiaohan Ding、Lance T. Wilhelm、Sophia Stil、Kirk McInnis Knutsen 和 Eugenia H. Rho。2025年。Conversate：通过交互式模拟与对话式反馈支持面试练习中的反思性学习。*Proc. ACM Hum.-Comput. Interact.* 9, 1, 文章GROUP9 (2025年1月), 32页。doi:10.1145/3701188[11] K Anders Ericsson 和 Herbert A Simon。1980年。作为数据的口头报告。*Psycholog- ical review* 87, 3 (1980), 215。[12] Jonathan St BT Evans。2008年。推理、判断与社会认知的双加工解释。*Annu. Rev. Psychol.* 59, 1 (2008), 255–278。[13] John H Flavell。1979年。元认知与认知监控：认知发展研究的新领域。*American psychologist* 34, 10 (1979), 906。doi:10.1037/0003-066X.34.10.906[14] Linda Flower 和 John R. Hayes。1981年。写作认知过程理论。*College Composition and Communication* 32, 4 (1981), 365–387。http://www.jstor.org/stable/ 356600 [15] Michael Gerlich。2025年。社会中的人工智能工具：对认知卸载的影响及批判性思维的未来。《社会》 15, 1 2025 6 10 3390 15010006 （），。doi: 。/soc Stephen T Homer。概念图绘制的比较分析：人类参与者与ChatGPT。《质量与数量》 59（2025年），4873–4892。doi:10.1007/ s11135-025-02211-w[17] Shih-Hong Huang、Ya-Fang Lin、Zeyu He、Chieh-Yang Huang和Ting-Hao Kenneth Huang。2024年。对话长度如何影响用户满意度？一项关于与大语言模型驱动的聊天机器人进行长度受控对话的案例研究。载于*CHI*计算机系统人为因素会议扩展摘要 （美国夏威夷州檀香山） *(CHI EA '24)*。美国计算机协会，美国纽约州纽约，文章编号 188, 13 10 1145 3613905 3650823 页。doi: 。/ 。Peiling [18] Jiang、Jude Rayan、Steven P. Dow和Haijun Xia。Graphologue：借助交互式图表探索大语言模型响应。载于第*36*届*ACM*用户界面软件与技术研讨会年度论文集 （美国加利福尼亚州旧金山） *(UIST '23)*。美国计算机协会，美国纽约州纽约，文章编号 3, 20 10 1145 3586183 3606737 页。doi: 。/ 。[19] Daniel Kahneman。《思考，快与慢》。Farrar, Straus and Giroux，纽约州纽约。[20] David Kirsh。2010。借助外部表征进行思考。《人工智能与社会》 25，4（2010），441–454。[21] Mina Lee、Katy Ilonka Gero、John Joon Young Chung、Simon Buckingham Shum、Vipul Raheja、Hua Shen、Subhashini Venugopalan、Thiemo Wambsganss、David Zhou、Emad A. Alghamdi、Tal August、Avinash Bhat、Madiha Zahrah Choksi、Senjuti Dutta、Jin L.C. Guo、Md Naimul Hoque、Yewon Kim、Simon Knight、Seyed Parsa Neshaei、Antonette Shibani、Disha Shrivastava、Lila Shroff、Agnia Sergeyuk、Jessi Stark、Sarah Sterman、Sitong Wang、Antoine Bosselut、Daniel Buschek、Joseph Chee Chang、Sherol Chen、Max Kreminski、Joonsuk Park、Roy Pea、Eugenia Ha Rim Rho、Zejiang Shen和Pao Siangliulue。2024年。智能交互式写作助手的设计空间。载于*2024*年*CHI*计算机系统人为因素会议论文集（美国夏威夷州檀香山） *(CHI '24)*。美国计算机协会，美国纽约州纽约，文章编号 1054, 35 10 1145 3613904 3642697 页。doi: 。/ 。Q. Vera Liao和Jennifer [22] Wortman Vaughan。2023。大语言模型时代的人工智能透明度：以人为本的研究路线图。arXiv:2306.01941 [cs.HC] https://arxiv.org/abs/ 2306 01941 。[23] David Chuan-En Lin、Fabian Caba Heilbron、Joon-Young Lee、Oliver Wang和Nikolas Martelaro。2024年。VideoMap：支持潜在空间中的视频探索、头脑风暴与原型设计。载于第*16*届创造力与认知会议论文集 （美国伊利诺伊州芝加哥） *(C&C '24)*。美国计算机协会，美国纽约州纽约， 311 327 10 1145 3635636 3656192 – 。doi: 。/ 。Jessy [24] Lin、Nicholas Tomlin、Jacob Andreas和Jason Eisner。面向决策的对话以实现人机协作。《计算语言学协会汇刊》 12 2024 892 911 10 1162 _ _00679 （），– 。doi: 。/ [25] tacl a Susan Lin、Jeremy Warner、J.D. Zamfirescu-Pereira、Matthew G Lee、Sauhard Jain、Shanqing Cai、Piyawat Lertvittayakumjorn、Michael Xuelin Huang、Shumin Zhai、Bjoern Hartmann和Can Liu。2024年。Rambler：通过大语言模型辅助的要点操控，以语音支持写作。载于*2024*年*CHI*计算机系统人为因素会议论文集 （美国夏威夷州檀香山） *(CHI '24)*。美国计算机协会，美国纽约州纽约，文章编号1043，19页。doi:10. 1145/3613904.3642217[26] Pauline Malaguti、Alexander J. Karran、Di Le、Hayley Mortin、Constantinos K. Coursaris、Sylvain Sénécal和Pierre-Majorique Léger。2025年。探究生成式人工智能中的交互摩擦：改善用户体验与决策。载于*SIGHCI 2024*年论文集。信息系统协会（AIS），泰国曼谷， 26 2024 26 。[27] 1990 https://aisel.aisnet.org/sighci / Thomas O Nelson。元记忆：理论框架与新发现。载于《学习与动机心理学》。第26卷。爱思唯尔，加利福尼亚州圣迭戈，125–173。doi: 10 1016 0079 7421 08 60053 5 。/S - ( ) - Joseph D. Novak和Alberto J. Cañas。[28] 2008年。概念图的基础理论及其构建与使用方法。技术报告 IHMC Cmap-Tools 2006-01 Rev 01-2008。佛罗里达州人类与机器认知研究所，佛罗里达州彭萨科拉。第1–31页。http://cmap.ihmc.us/Publications/ResearchPapers/ TheoryUnderlyingConceptMaps.pdf[29] Peter Pirolli和Stuart Card。2005。意义建构过程以及通过认知任务分析识别的分析人员技术杠杆点。载于国际情报分析会议论文集，第5卷。MITRE公司，美国弗吉尼亚州麦克莱恩，2–4。[30] Kevin Pu、K. J. Kevin Feng、Tovi Grossman、Tom Hope、Bhavana Dalvi Mishra、Matt Latzke、Jonathan Bragg、Joseph Chee Chang和Pao Siangliulue。2025年。IdeaSynth：通过不断演化并组合构想要素，并结合基于文献的反馈，迭代开展研究构想开发。载于*2025*年*CHI*计算机系统人为因素会议论文集（*CHI '25*）。美国计算机协会，美国纽约州纽约，文章编号145，31页。doi:10.1145/3706598.3714057[31] Leon Reicherts、Zelun Tony Zhang、Elisabeth von Oswald、Yuanting Liu、Yvonne Rogers和Mariam Hassib。2025年。人工智能，帮我思考——但要为我自己思考：通过提供不同类型的认知

支持。载于2025年CHI计算机系统人为因素会议论文集（CHI '25）。美国计算机协会，美国纽约州纽约，文章编号 255, 19 10 1145 3706598 3713295 页。doi: 。/ 。[32] Valerie F Reyna和Charles J Brainerd。模糊痕迹理论：阶段性综合。学习与个体差异 7，1（1995），1–75。[33] Yvonne Rogers。2025年。为何在借助生成式人工智能学习时值得付出努力。《国际艺术、人文与社会科学期刊》06，特刊（8月 2025 1 6 10 56734 6 1），– 。doi: 。/ijahss.v nsa Sherry [34] Ruan、Jacob O. Wobbrock、Kenny Liou、Andrew Ng和James A. Landay。2018。触屏手机上两种语言短消息的语音与键盘文本输入比较。《ACM交互、移动、可穿戴与普适技术汇刊》 1, 4, 159 2018 23 10 1145 3161187 文章编号（1月），页。doi: 。/ Mike Scaife和Yvonne Rogers。外部认知：图形表征如何发挥作用？国际人机交互研究期刊 45，2（ 1996 185 213 10 1006 1996 0048），– 。doi: 。/ijhc. 。[35] Daniele Schicchi、Carla Limongelli、Vito Monteleone和Davide Taibi。2026。深入审视ChatGPT在教育概念图生成中的作用。交互式学习环境 34, 1 2026 276 296 10 1080 10494820 2025 2497110（），– 。[36] doi: 。/ 。。Donald A. Schön。《反思性实践者：专业人士在行动中如何思考》。Basic Books，纽约。[37] Anjali Singh、Zhitong Guan和Soo Young Rieh。2025。借助元认知提示增强生成式人工智能搜索中的批判性思维。arXiv: 2505 24014 [ ] 2505 24014 。cs.HC https://arxiv.org/abs/ 。[38] Chandra Sripada和Aman Taxali。意识流中的结构：来自言语化思维协议与自动化文本分析方法的证据。意识与认知 85（2020），103007。[39] Sangho Suh、Meng Chen、Bryan Min、Toby Jia-Jun Li和Haijun Xia。2024年。Luminate：利用大语言模型进行设计空间的结构化生成与探索，以支持人机协同创作。载于2024年CHI计算机系统人为因素会议论文集（美国夏威夷州檀香山）（CHI '24 644, 26 10 1145 3613904）。美国计算机协会，美国纽约州纽约，文章编号 页。doi: 。/ 。[40] Sangho Suh、Bryan Min、Srishti Palani和Haijun Xia。2023。Sensecape：利用大语言模型实现多层次探索与意义建构。载于第36届ACM用户界面软件与技术研讨会年度论文集（美国加利福尼亚州旧金山）（UIST '23）。美国计算机协会，美国纽约州纽约，文章编号 1, 18 10 1145 3586183 3606756 页。doi: 。/ 。[41] John Sweller。问题解决过程中的认知负荷：对学习的影响。认知科学 12，2（1988），257–285。[42] Lev Tankelevitch、Elena L. Glassman、Jessica He、Majeed Kazemitabaar、Aniket Kittur、Mina Lee、Srishti Palani、Advait Sarkar、Gonzalo Ramos、Yvonne Rogers和Hari Subramonyam。2025年。思维工具：面向利用生成式人工智能理解、保护并增强人类认知的研究与设计。载于CHI计算机系统人为因素会议扩展摘要论文集（CHI EA '25）。美国计算机协会，美国纽约州纽约，文章编号 804, 8 10 1145 3706599 3706745 页。doi: 。/ 。[43] Lev Tankelevitch、Elena L. Glassman、Jessica He、Aniket Kittur、Mina Lee、Srishti Palani、Advait Sarkar、Gonzalo Ramos、Yvonne Rogers和Hari Subramonyam。2025年。利用生成式人工智能理解、保护并增强人类认知：2025年CHI思维工具研讨会综述。arXiv:2508.21036 [cs.HC] https://arxiv.org/abs/2508.21036[44] Lev Tankelevitch、Viktor Kewenig、Auste Simkute、Ava Elizabeth Scott、Advait Sarkar、Abigail Sellen和Sean Rintel。2024年。生成式人工智能的元认知要求与机遇。载于2024年CHI计算机系统人为因素会议论文集（美国夏威夷州檀香山）（CHI '24）。美国计算机协会，美国纽约州纽约，文章编号680，24页。doi:10. 1145/3613904.3642902[45] Nadine Wagener、Leon Reicherts、Nima Zargham、Natalia Bartłomiejczyk、Ava Elizabeth Scott、Katherine Wang、Marit Bentvelzen、Evropi Stefanidi、Thomas Mildner、Yvonne Rogers和Jasmin Niess。2023。SelVReflect：一项引导式虚拟现实体验，促进对个人挑战的反思。载于2023年CHI计算机系统人为因素会议论文集（德国汉堡）（CHI '23）。美国计算机协会，美国纽约州纽约，文章编号 323, 17 10 1145 3544548 3580763 页。doi: 。/ 。[46] Cui-Yu Wang 、 Bao-Lian Gao 和 Shu-Jie Chen。2023。项目式学习环境中的元认知支架对学生元认知能力与计算思维的影响。教育与信息技术 29，5（2023年7月），5485–5508。doi:10.1007/s10639-023-12022-x [47] Karl E. Weick。1995。组织中的意义建构。组织科学基础，第3卷。Sage Publications，加利福尼亚州千橡市。[48] Yunhao Xing、Jerrick Ban、Timothy D Hubbard、Michael Villano和Diego Gómez-Zará。2025年。沉浸于我的想法：利用虚拟现实与大语言模型可视化用户的想法与思维。载于第30届智能用户界面国际会议配套论文集（IUI '25 Companion）。美国计算机协会，美国纽约州纽约， 60 65 10 1145 3708557 3716330 – 。doi: 。/ 。[49] Koji Yatani、Zefan Sramek和Chi-Lan Yang。人工智能作为超越启发式：在人机交互中培养高阶思维技能。arXiv:2409.09218 [cs.HC] https://arxiv.org/abs/2409.09218 [50] Rui Zhang、Ziyao Zhang、Fengliang Zhu、Jiajie Zhou 和 Anyi Rao。2024年。Mindalogue：大语言模型驱动的非线性交互，助力有效学习与任务探索。arXiv: 2410 10570 [ ] 2410 10570 。cs.HC https://arxiv.org/abs/。[51] Xiaoyu Zhang、Senthil Chandrasegaran 和 Kwan-Liu Ma。。ConceptScope：基于领域本体组织和可视化文档中的知识。载于*2021*年*CHI*计算机系统人为因素会议论文集（日本横滨）（*CHI '21*）。美国计算机协会，纽约州纽约，美国，文章 19, 13 10 1145 3411764 3445396 页。数字对象标识符：。/。[52] Xiaoyu Zhang、Jianping Li、Po-Wei Chi、Senthil Chandrasegaran 和 Kwan-Liu Ma。2023年。ConceptEVA：基于概念的文档摘要交互式探索与定制。载于*2023*年*CHI*计算机系统人为因素会议论文集（德国汉堡）（*CHI '23*）。美国计算机协会，纽约州纽约，美国，文章204，16页。数字对象标识符：10.1145/3544548.3581260[53] Zheng Zhang、Jie Gao、Ranjodh Singh Dhaliwal 和 Toby Jia-Jun Li。2023年。VISAR：结合可视化编程与快速草稿原型设计的人机协作论证写作助手。载于第*36*届*ACM*用户界面软件与技术研讨会论文集（美国加利福尼亚州旧金山）（*UIST '23*）。美国计算机协会，纽约州纽约，美国，文章5，30页。数字对象标识符：10.1145/3586183.3606800[54] Zheng Zhang、Weirui Peng、Xinyue Chen、Luke Cao 和 Toby Jia-Jun Li。2025年。LADICA：面向共址团队协作中生成式人工智能认知辅助的大型共享显示界面。载于*2025*年*CHI*计算机系统人为因素会议论文集（*CHI '25*）。美国计算机协会，纽约州纽约，美国，文章147，22页。数字对象标识符：10.1145/3706598.3713289

A Pirolli 和 Card模型

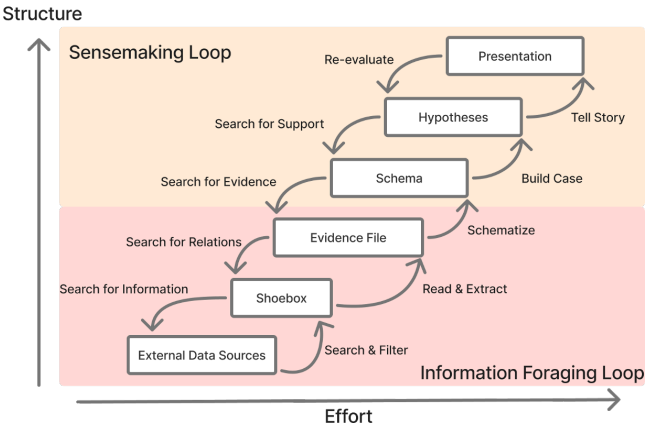

图14：Pirolli 和 Card 提出的原始模型 [29]。

B 形成性研究中言语化思维的转录文本示例

P2讲述了一个研究项目，该项目使用增强现实（AR）音乐游戏将家务游戏化。这段独白记录了贯穿项目多个阶段的零散想法。

“嗯，最近我读了几篇论文，主要是关于挑战的……然后我发现，嗯，他们其实没办法享受其中，也难以集中注意力……所以我在考虑……用游戏，尤其是增强现实游戏，来帮助用户……现在，我的导师似乎觉得这个方向很有意思，但我目前的主要问题还是，呃，研究问题……首先，嗯，该怎么说呢？……可能有三个研究问题……呃，还有另一个研究问题，是什么来着？我有点忘了……呃，现在我的想法是……我需要先做一个原型……我发现音乐游戏是个挺不错的选择……它没有任何暴力元素……只是跟着音乐，呃，按节拍点击……还有……音乐游戏可以让用户活动身体和双手……我在考虑设计一个增强现实场景，让用户一边玩游戏一边擦窗户……总之，这份提案是课程作业……我需要规划时间线……我想现在可以先写下，呃，我打算如何设计这个游戏……节拍点……能量条……呃，我觉得在需要点击的时刻到来之前，应该有一个动画……嗯，我想我可以这样考虑节拍点……我需要控制难度……不能设计得太难，否则用户可能会太累……呃，我之前读到过关于难度控制的内容……嗯……最后，我该如何开展实验？我需要让用户填写两份问卷，一份关于专注度，一份关于愉悦感……客观数据很重要……呃，比如擦窗前后的干净程度，但这很难量化……还有用户完成任务所花的时间……感觉时间挺紧的。不过我们现在团队里有几个人，也许可以一起做……好，我想大概就是这些了。”

C 评估研究的额外细节

C.1 基线对话片段示例

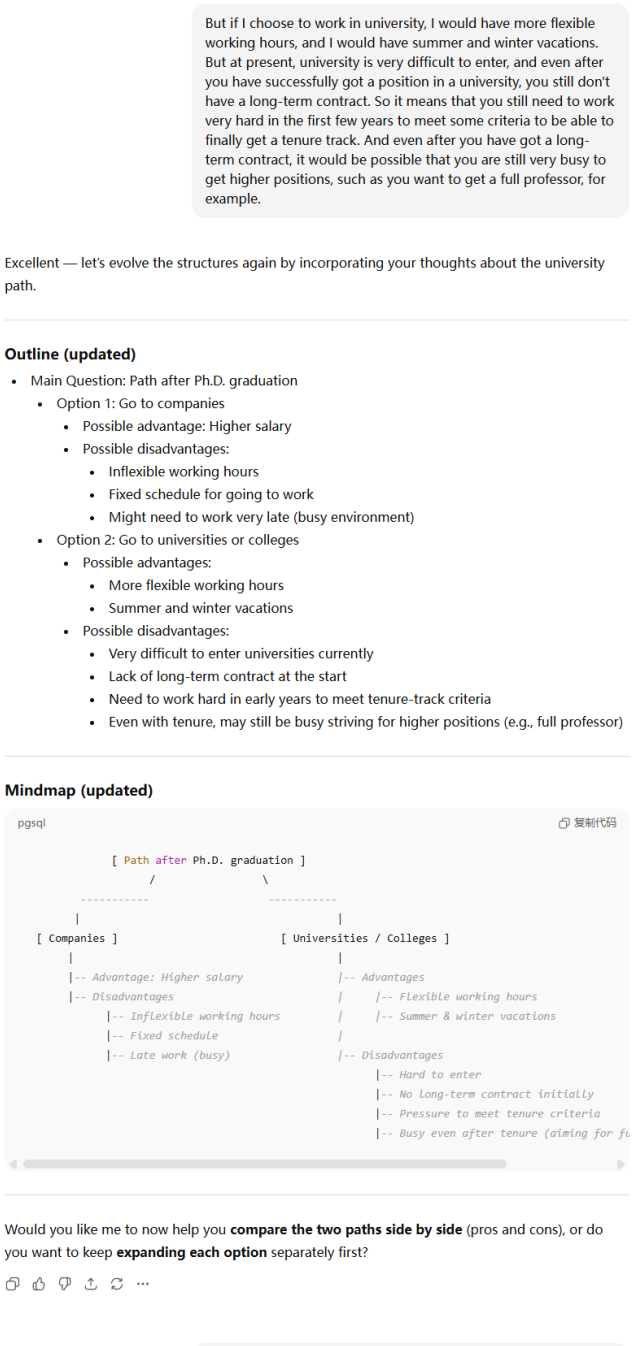

图15：参与者基线任务试次的摘录示例。

C.2 参与者人口统计

表1：参与者的人口统计、口述和生成式人工智能使用情况

参与者	出席情况	学位	领域	听写使用情况	生成式人工智能使用情况
P1	在线	硕士	HCI	每年	专家
P2	线下	本科	设计	每月	专家
P3	在线	硕士	健康信息	每月	专家
P4	线下	硕士	HCI	日常	普通用户
P5	线下	硕士	HCI	每周	休闲用户
P6	线下	硕士	HCI	每月	普通用户
P7	线下	硕士	计算机科学	每周	普通用户
P8	线下	硕士	HCI	日常	专家
P9	在线	硕士	工业设计	日常	休闲用户
P10	在线	硕士	计算机科学	每年	专家
P11	在线	硕士	HCI	每年	普通用户
P12	线下	本科	游戏设计	从未	普通用户

C.3 评估研究的参与者主题

评估研究中每位参与者的主题和详细信息列于表2和3。为保护参与者隐私，具体地点、独特的研究项目描述和个人商业创意等身份识别信息均已匿名化或泛化处理。

表2：口语条件下参与者的主题

参与者	口语条件主题	任务类别
P1	如何成为软件架构师	策略制定
P2	为用户体验研究员实习面试准备自我介绍	策略制定
P3	谈论一项进行中的项目及其当前挑战	项目规划
P4	完成海外博士学位后的职业发展选择	比较决策
P5	规划如何为一项研究项目选择分类框架	项目规划
P6	比较两个主要城市的生活质量和职业前景	比较决策
P7	开发海报生成应用并撰写报告	项目规划
P8	总结三个项目并规划未来工作	项目规划
P9	谈论行业环境中的一个进行中的项目	项目规划
P10	权衡找工作或攻读博士学位的选择	比较决策
P11	思考即将到来的学期的学习与生活计划	策略制定
P12	个人成长和五年计划	策略制定

表3：基线条件下参与者的主题

参与者	基线条件主题	任务类别
P1	个人发展：在A市与B市之间做出选择	比较决策
P2	规划作为国际学生如何省钱	策略制定
P3	找出难以入睡背后的原因以及如何解决	策略制定
P4	为指定的人机交互主题概述研究路线图	项目规划
P5	规划多人多日旅行行程，涵盖后勤安排	策略制定
P6	在获得博士学位后，于学术生涯与业界职位之间做出抉择	比较决策
P7	组织一场关于青少年教育方法的辩论	策略制定
P8	决定是寻求就业机会，还是攻读博士后职位	比较决策
P9	谈论一项正在进行的学术研究项目	项目规划
P10	描述进行中的项目并概述下一步	项目规划
P11	描述进行中的项目并安排未来两周的任务	项目规划
P12	制定盈利计划，邀请他人成为商业伙伴	策略制定

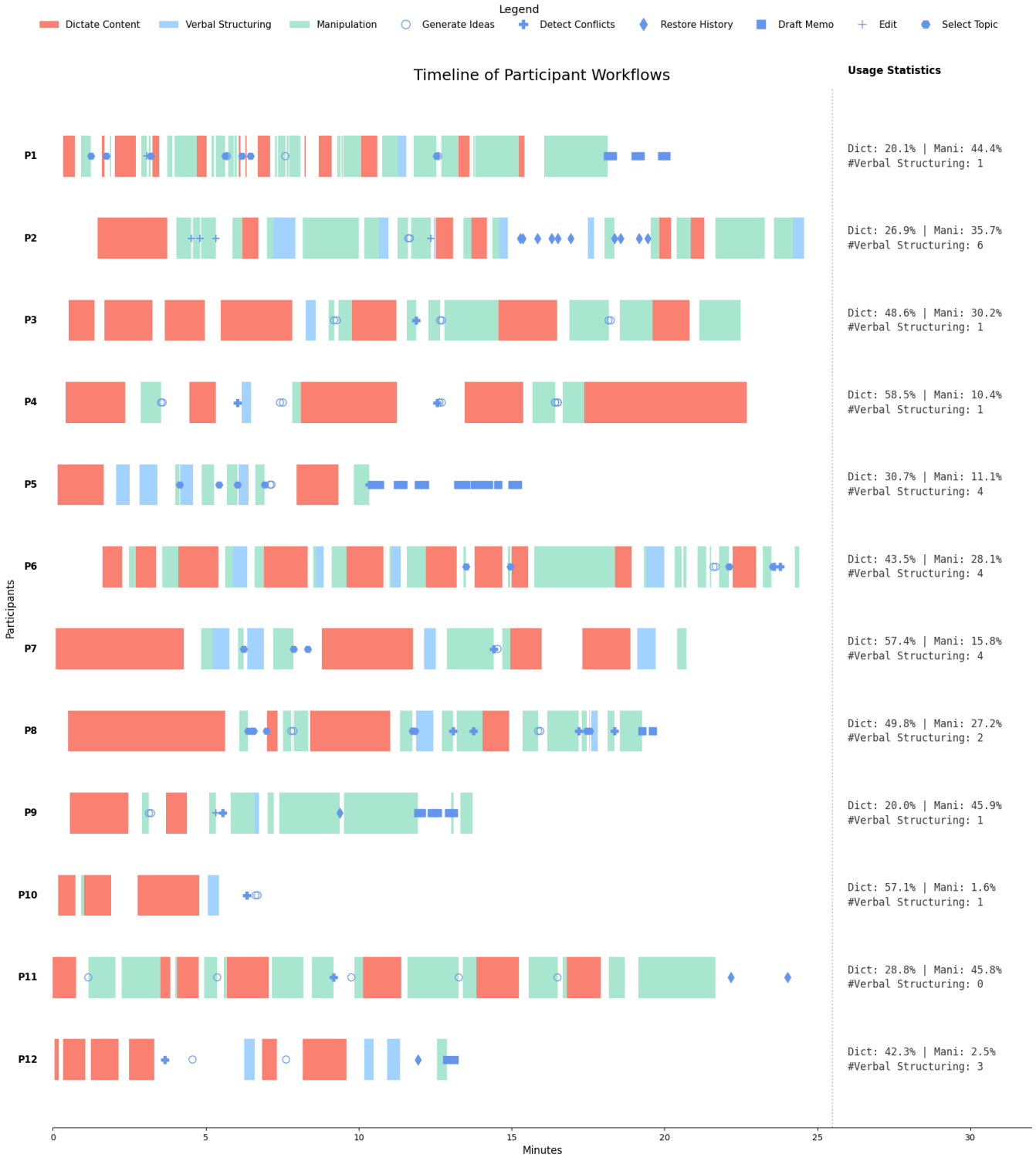

图16：可视化每位参与者各项任务工作流程的时间线。每次使用各项功能均标注为单个点。语音讲述以及对节点的操作标注为时间段。每位参与者的使用统计显示在右侧，包括口述内容（Dict）和操作（Mani）的时间百分比，以及口头结构化的次数（# Verbal Structuring）。时间百分比按该活动在整个会话中所占的总时间计算。

D 参与者工作流程

图16展示了 Oral-ity 在实验试验时间线上的功能使用情况。这里我们介绍每位参与者采用的整体工作流程和策略。

• P1 的工作流程高度迭代且注重动手操作，以手动整理为核心。他们首先通过简短、专注的语音输入，将各项技能和职业目标逐一添加到画布中。每次添加少量内容后，他们都会花大量时间手动排列节点，利用空间布局建立联系并梳理思路。这种少说一些、再花大量时间手动整理的模式，是他们主要的交互方式。直到最后，他们才通过语音指令启用人工智能，将手动创建的大型集群拆分为更细致的子主题；人工智能在此只是最后的“清理”步骤，而非主要的组织工具。

• P2 将该工具视为叙事助手，采用几乎完全由语言驱动、很少进行手动操作的工作流程。他们首先口述自己的学术与职业经历。随后，他们通过一系列精准的文本和语音指令，让人工智能将这些事实整理成清晰的类别（例如“研究助理经历”和“专业工作经历”）。值得注意的是，他们还利用人工智能回答问题，请它综合自己的经历，并说明自己为何适合目标职位。他们使用“思维演进”功能，体现出一种迭代式的对话方法：尝试不同的指令，以找到效果最佳的人工智能生成结构和摘要。

• P3采用了问题解决叙事工作流程，主要依靠语音输入，很少手动排列节点。他们首先口头阐明了一个明确的问题（一篇被拒的学术论文）及其背景。随后，他们通过进一步的语音输入，逐步拆解具体的审稿人反馈并开展头脑风暴，提出可能的解决方案，从而加深理解。他们主要通过谈论问题空间来进行互动，让人工智能在信息生成的过程中加以整理。他们使用语音指令进行高层次重组，以“让它看起来更清晰”，但这只发生在他们已经通过语音充分外化了思维过程之后。

• P4采用了“漏斗式”策略来聚焦。他们首先进行了一次全面的“头脑倾倒”，通过语音外化当前生活和职业不确定性的方方面面。经过短暂的手动排列后，他们使用一条关键语音指令，引导人工智能精简画布，要求它隐藏与签证事宜相关的主题，只聚焦于他们“毕业后的不确定性”。利用人工智能进行范围界定和管理注意力，是其工作流程的核心。画布聚焦于核心问题后，他们便开始对该主题进行深入、细致的口头探索，依靠工具记录详尽的想法，无需再进行大量操作。

• P5 的工作流程以纠正性、层级化的指令对话为特征，几乎不涉及手动操作。在最初进行口头思维倾倒后，他们使用一系列精确的语音命令来指导人工智能整理笔记。当人工智能第一次尝试重组时不太符合要求，用户没有转而进行手动修改；相反，他们立即发出更清晰、更具体的语音指令，从而有效地细化指令，以获得预期结果。他们在更细致的层面上重复了这一模式，选定单个主题，并再次发出指令，仅重构其内容。这展现了一种复杂的、由语言驱动的方法：用户将人工智能视为组织助手，可以通过迭代式口头指令对其进行引导。

• P6 采用增量式方式构建论证，利用人工智能逐步进行分类。与一开始就进行大量头脑倾倒的用户不同，这位用户以一连串简短而专注的语音片段表达想法，每次添加一两个相关想法（例如一个优点和一个缺点）。在补充完一条推理思路中的要点后，他们会发出简单的语音指令，例如“将第一个理由整理到一个新的子主题中”，让人工智能把这些要点归为一组。这种逐步补充内容、再下达指令进行小幅整理的过程不断重复，直至完成比较分析。这一工作流程也涉及大量手动操作。

• P7 展现了一种高度结构化、由语言驱动的工作流程，几乎没有进行手动节点操作。在最初进行全面的“头脑倾倒”后，他们立即使用一系列具体的语音指令来构建画布。他们的过程呈现出明显的层级性：首先，他们下达指令，将所有内容自上而下地重组为一个分析框架（例如“背景”“挑战”“方法”）。接着，他们“缩小范围”，选中一个新的部分，并下达另一条细致指令来重构其内部布局。这种在宏观和微观组织中都使用精准语音指令的模式，体现出一种策略：不采用手动排列，而是通过口头指挥人工智能构建复杂的多层结构。

• P8进行了一段问题解决叙事，使用该工具时几乎不依赖手动排列。他们首先描述了彼此脱节的项目以及一种核心的困惑感。他们没有手动移动节点来寻找联系，而是用语音直接向画布求助，并借助AI生成的问题引导自己的反思。这带来了一次关键的概念性突破。在获得这一洞见之后，其交互仍以语言为基础；他们发出语音命令，让人工智能在新主题下综合其项目，并自动“修复”其推理中的冲突。他们的工作流程并非在于以视觉方式组织节点，而是通过对话式交互和基于命令的交互来解决一个概念问题。

• P9 主要将该工具视为可视化项目规划画布，主要依靠手动整理，同时避免人工智能驱动的重组。他们先进行全面的“头脑倾倒”，将所有项目组成部分外化。随后，他们的主要活动是细致地手动排列节点，以构建可视化计划。值得注意的是，在尝试使用语音指令进行重组时，他们发现

人工智能简化后的输出不尽如人意，于是立即恢复到自己偏好的手动整理布局。他们的工作流程表明，他们倾向于使用语音进行初步内容捕捉，并在后续所有整理工作中采用手动操作。

• P10采用了条理清晰的“先说出来，之后再整理”工作流程，几乎不进行手动节点操作。他们首先通过口头表达，将决策问题的各个方面都外化出来——列出优点、缺点和生活方式方面的考量。想法呈现在画布上后，他们没有手动排列，而是用一句简洁明确的语音指令（“让图表更清晰”），让人工智能一次性整合并构建整个画布的结构。他们的策略是先全面开展口头头脑风暴，再通过人工智能一次性果断地重组内容。

• P11采用了均衡、迭代的多模态策略。他们先用语音概述大致内容，随后进入一个固定循环：先通过语音添加新一层细节，然后立即进行大量手动操作，拖动并重新排列新节点，将其直观地整合起来。这种先说后手动整理的紧密循环，是他们逐步梳理想法的主要方法。进行重大重组时，他们会使用高层级指令，让人工智能将整个画布重新整理为更具主题性的摘要。

• P12进行了高度反思性的辩证过程，几乎完全由语言驱动，对节点的手动操作极少。在外化其最初的想法后，他们依靠一系列逐渐更为复杂的语音指令来指导人工智能。他们命令它不仅重组这些想法，还要进行分析，将想法分类为“重叠”和“冲突”要点。这表明了一种工作流程：用户将人工智能视为分析伙伴，使用口语同时进行内容生成和复杂的结构转换，从而在很大程度上绕过了手动排列。

E 提示词

E.1 基线的默认提示词

在使用基线开展思维组织任务之初，会向 ChatGPT 发送以下提示词，以便使其可与 Orality 进行比较。

我想通过一些图形和文字信息逐步组织我的想法。我将主要通过口述来输入我的想法。我希望你模拟一个能够执行以下功能的工具：*1.* 你可以提供多层次结构，以表示我刚刚输入的内容。该结构可以是思维导图、大纲，或两者兼有。思维导图不应与大纲相似；你应使用文字和虚线来模拟图形化思维导图。*2.* 每当我输入新内容时，你应将新内容添加到现有内容中，从而使整体内容不断“演进”。

当你准备好后，我就开始。

E.2 Orality 中使用的提示词

任务	提示模板
提取主题和内容节点	你是某位发言者的听众。请分析给定文本并提取关键信息，以短句实体的形式呈现，而非关键词或短语。如果主语是说话者本人，则无需添加主语。某些句子实体可以归入同一个“观点”，即说话者的陈述、判断、观点或相关事项等。如果用户有一些指令，请遵循这些指令来创建“观点”。在每个“观点”下生成一到三个句子实体，使得每个独立实体都包含完整信息，并避免两个实体包含相似信息。不要生成内容相似的实体。请按以下格式将它们放入列表：[“主题”：……，“实体”： [……，……，……] ,……]
重组指令	你是一名内容整理器。请分析给定的大纲，并根据用户的具体指令重新整理其中的元素。在当前大纲中，有一些主题和属于这些主题的实体。你需要按照指令生成新的大纲，并挑选现有实体将它们归入新大纲，或者从用户的完整文本中提取相关但当前不存在的实体。这里有三种常见的与用户全文相关但并非现有的实体。可能存在三种常见的指令示例：1. 用户要求采用一种新结构并指定具体主题时，你应使用用户定义的主题作为输出中的主题。2. 用户要求采用一种新结构但未指定新主题时，你应提出与该结构相关的主题。3. 用户要求基于当前结构添加新主题时，你应添加这些将新主题添加到现有大纲，并从全文中提取新的相关内容实体。如果用户的询问不属于这些情况，你应尽可能谨慎地处理，并且尽可能保留现有结构。按以下格式输出新大纲： ["topic": ..., "entities": ["type": 如果这是现有实体，请勿更改其类型；否则，请设置 to 1 "text": ……] ,……]
问我问题-问题	你是一位深思熟虑的对话引导者，帮助用户整理思路。给定一个主题及其当前内容，生成1-2个引导性问题，以帮助用户探索更深入地探讨该主题，或补充缺失的信息。这些问题应当：1. 采用开放式，并能引发思考 2. 帮助用户通过具体细节、示例或见解扩展主题 3. 与主题中的现有内容直接相关 4. 鼓励用户分享经历、具体示例或深入分析 5. 以有帮助的提示词形式表述（例如：“有哪些具体挑战……”、“你能描述一下……”、“你是如何做到的……”） 6. 避免使用是非问题 7. 简洁明了，但要具体 将回复格式设为简单的 JSON 数组，数组元素为问题字符串： ["问题 1？ ", "问题 2？ "] 每个主题生成 1–2 个问题（内容很少的主题生成 2 个问题，有一些内容的主题生成 1 个问题）。

任务	提示模板
查找冲突	你擅长识别想法和思维中的逻辑冲突与矛盾。 分析成对的内容，找出以下冲突： 1. 直接矛盾——彼此直接对立的陈述 2. 逻辑不一致——不可能同时为真的陈述 3. 价值冲突——相互对立的价值观、优先事项或原则 4. 策略冲突——针对同一问题彼此矛盾的方法 5. 假设冲突——彼此不相容的潜在假设 对于每一对，请判断： - 是否存在有意义的冲突？（而不仅仅是主题或视角不同） - 这属于哪种类型的冲突？ - 你的确信程度有多高？（1-10 分） - 简要说明为何这是冲突 仅以 JSON 数组作答（不要使用 Markdown 格式），其中每个元素代表一个检测到冲突： ["pair_id": 0, "has_conflict": true, "conflict_type": "直接矛盾", "confidence": 8, "reason": "一方称 X 为真，而另一方称 X 为假"] 仅包含你检测到真正冲突的配对（has_conflict: true）。请有选择性地 - 同一主题上的不同视角并不一定构成冲突。聚焦于实质性 如果两者都为真，就会造成逻辑问题的矛盾。如果 [] 未发现冲突，则返回空数组。
导出报告	请根据以下整理好的内容和所选 风格：'comprehensive'：你正在帮助用户整理和综合自己的想法。 创建一份全面的个人总结，其中： 1. 综合提炼其内容中的关键主题和见解 2. 识别其思考中的重要模式和联系 3. 突出重要见解和潜在的后续步骤 4. 采用个人化、反思性的语气（直接与用户对话） 5. 将思路整理成清晰、合乎逻辑的部分 结构如下：关键主题、重要见解、联系与模式，以及后续步骤。 'executive'：为用户的想法撰写一份高层次的个人总结。重点关注： 1. 其内容中最重要的主题和见解 2. 他们可能需要考虑的关键决策或行动 3. 其思考所带来的战略影响 4. 使用个人化、直接的语言（使用“你”和“你的”） 5. 简洁而全面的概述 简明扼要地概括，但要抓住其思考中的核心见解。 'bullet'：使用清晰的项目符号创建结构化的个人总结： 1. 将用户的想法整理成逻辑分组 2. 使用清晰易读的项目符号 3. 在适当时列出行动事项或后续步骤 4. 使用个人化的语言（直接称呼用户） 5. 让内容便于快速浏览和日后查阅 撰写一份连贯、结构清晰的备忘录，综合这些主题并提供有价值的见解。备忘录应内容详尽但易于阅读，突出关键主题以及主题之间的关系并提供可执行洞见。