arXiv:2611.31899v1 [cs.HC] 24 Nov 2026

超越平均收益：探究结构化视觉人工智能学习中的公平—脆弱性张力

MINGYUAN LI，宾夕法尼亚大学，美国DUAN WANG，加利福尼亚大学圣迭戈分校，美国ERICK PURWANTO，西交利物浦大学，中国YIHONG WANG*，
西交利物浦大学，中国

开放式对话式人工智能要求学习者提出问题、组织学习路径，并跟踪不断变化的问题状态。结构化视觉界面可能会减轻这些要求，但更高的平均评价并不能证明不同背景和设备的学习者都能一致地获得这种支持。我们开展了一项关于图算法学习的随机研究，将结合预定义学习阶段、对话式解释和动态算法可视化的 VisualCodeChat（VCC）与标准的基于文本的聊天机器人进行了比较。我们分析了来自大学生的282份任务后问卷。在未经调整的比较中，VCC获得了显著更高的可用性和感知有效性评价。我们未发现两种条件在效率、满意度或使用后情绪方面存在显著差异。探索性条件内分析发现了与既往聊天机器人使用经历相关的差异：在标准人工智能条件下，低频用户报告了较低的采纳度和学习管理评价；而不舒适的既往聊天机器人使用经历则与VCC条件下较低的交互和布局评价相关。这些发现表明，改善平均交互体验只是让人工智能支持的学习更具包容性的一部分。未来的评价应考察具有不同既往经验的学习者是否能够在不同设备上获得指导、克服困难，并独立展现理解。

CCS 概念：• **以人为中心的计算** → **人机交互实证研究**；• **应用计算** → 交互式学习环境；•**社会与专业议题** →计算教育。

附加关键词和短语：教育聊天机器人、视觉支架、情感舒适度、教育公平、计算教育

1 引言

人工智能支持的学习系统日益为学生提供按需解释、指导和反馈。学习者面临的交互要求会影响这类支持的可及性。这个问题在
图算法学习中尤为突出：学生需要推理抽象结构，同时追踪算法状态的变化。算法可视化可以让这些过程更加直观，并支持概念理解[21, 40], 而对话式人工智能则允许学生使用自然语言请求解释和指导 [25]。然而，开放式交互要求学习者提出有用的问题、判断接下来该问什么，并将文本解释与不断变化的算法状态联系起来。这些要求可能使人工智能支持的学习效果部分取决于学习者对对话系统的既往熟悉程度。这引出了一个问题：平均评价良好的系统，能否为既往聊天机器人使用情况不同的学习者提供同样易于使用的支持?

结构化视觉人工智能为这一问题提供了一种可能的应对方式。与要求学习者完全通过自由形式提示来组织交互不同，结构化界面可以将学习划分为预定义阶段，

*通讯作者。

作者联系方式：Mingyuan Li，宾夕法尼亚大学，美国宾夕法尼亚州费城，myuli@engineering.upenn.edu；DuanWang，加利福尼亚大学圣迭戈分校，美国加利福尼亚州拉霍亚，duw003@ucsd.edu；Erick Purwanto，西交利物浦大学，中国苏州，Er-ick.Purwanto@xjtlu.edu.cn；Yihong Wang，西交利物浦大学，中国苏州，Yihong.Wang@xjtlu.edu.cn。

提供循序渐进的对话式指导，并将解释与算法执行的视觉表征同步起来。这类设计建立在长期积累的证据之上：恰当整合的言语和视觉表征能够支持学习 [36, 40]。通过将组织学习路径和追踪算法状态的部分工作从学习者转移到界面，结构化视觉人工智能或许能够减少与经验相关的、妨碍学习者使用人工智能支持的障碍。我们将交互公平性这一术语严格限定为：具有不同既往聊天机器人使用情况的学习者，是否报告了同样易用且具有支持性的交互。这一框架关注的是学习者感知到的交互可及性，而非更广泛的教育公平或学习成果平等。

然而，降低最初的交互障碍，并不意味着支持在所有学习者背景和使用条件下都同样有效。先前情感体验会影响学习者接近和理解技术中介学习的方式 [10, 44], ，而访问技术时所处的条件则可能限制其界面体验 [63]。结构化界面可能使学习路径更加清晰，但其固定的组织方式和视觉呈现，在学习者需要进一步澄清、回顾先前材料，或通过不同设备进行交互时，也可能缺乏足够的适应性。这些可能性表明，降低进入人工智能支持的学习交互的门槛，与在学习者参与之后持续提供有用支持之间存在区别。因此，我们将公平性—脆弱性关系视为一种潜在的评价张力：能够减少某一交互困难来源的设计特征，仍可能无法解决其他依赖于体验或情境的局限。我们考察较高的平均评价是否会与学习者既往经验和使用条件相关的差异并存。

既有研究为算法可视化的教学价值、情感在学习中的作用，以及技术情境的重要性提供了大量证据 [10, 21, 40, 44, 63]。然而，这些视角促使我们不仅考察人工智能学习界面是否获得更高的平均评价，还要考察这些评价如何分布在具有不同既往经验和设备的学习者之间。因此，我们提出两个问题。第一，与传统的基于文本的人工智能界面相比，结构化视觉人工智能在感知可用性、有效性、效率、满意度和使用后情绪方面有何差异？第二，在每种界面条件下，学习者的评价如何随先前的聊天机器人使用频率、与聊天机器人的先前情感体验以及设备类型而变化？

为探究这些问题，我们在图算法学习情境中开展了一项受控研究，并分析了282份任务后问卷。参与者被随机分配至VisualCodeChat（VCC）条件或标准人工智能条件，并完成相同的三个图算法学习任务序列。两种条件使用相同的大语言模型（LLM）和教学序列；其中，VCC结合了预定义的学习阶段、对话式解释和动态算法可视化，而标准人工智能条件则提供传统的基于文本的对话界面。完成任务后，参与者报告了对可用性、有效性、效率、满意度和情绪体验的感知。我们还考察了先前的聊天机器人使用频率、先前与聊天机器人的情感体验，以及设备类型等学习者和情境变量。这些测量反映的是自我报告的交互体验和感知到的学习支持，而不是经过独立评估的学习增益。

结果表明，结构化视觉人工智能带来的优势是有选择性的，而非全面的。使用 VCC 的参与者报告的可用性和感知有效性显著高于使用标准人工智能的参与者；而在效率、满意度或使用后情绪方面，我们未发现两种条件之间存在显著差异。进一步的探索性条件内分析揭示了被这些平均值比较所掩盖的差异。在标准人工智能条件下，不常使用聊天机器人的用户对用户采纳和学习管理的评价较低；而在 VCC 条件下，未报告相应的显著频率差异。在 VCC 条件下，之前有过

不愉快聊天机器人体验的参与者在若干交互相关维度上的评分较低；笔记本电脑用户对界面的评分则高于平板电脑或智能手机用户。这些子群体模式并不能证明 VCC 消除了与经验相关的差异，也不能证明其受先前情绪体验的影响更大；相反，它们表明，平均水平的提升可能与感知支持方面受背景和情境影响的差异并存。

本研究有三项贡献。实证方面，本研究刻画了结构化视觉人工智能获得更高评价的方面，尤其是在可用性和感知有效性方面。概念方面，本研究区分了降低与经验相关的进入人工智能支持交互的障碍，与在不同既往经验和使用条件下持续提供可比支持之间的差异。这一区分将公平—脆弱性张力界定为未来评价需要探讨的问题。最后，研究结果为设计方向提供了启示，例如可恢复的交互路径、请求本地化辅助的机制、跨设备评价，以及对学习成果的独立评估。

2 相关工作

2.1 教育人工智能与对公平的追求

教育人工智能可以为不同学习者提供量身定制的支架式支持。支架式支持建立在维果茨基的最近发展区概念之上 [62] ，描述指导如何帮助学习者完成他们尚无法独立完成的任务 [64]。智能辅导系统和聊天机器人是人工智能支持的支架式支持的常见形式[19]。这些技术可以在自主学习过程中提供个性化反馈。例如，人工智能驱动的语言学习工具可以提供可定制且低成本的练习机会 [50]。Son 等人 [55] 讨论了用于语言学习的文本和语音人工智能工具。学生表示，使用 ChatGPT 有助于他们进行外语学习，尤其是在写作、语法和词汇方面 [24]。人工智能辅助教学也广泛应用于计算教育。例如，使用 ChatGPT 支持编程教学可以提升学生的计算思维技能和学习动机 [67]。CodeAid 等工具使用大语言模型引导学生自行解决编程问题 [25]。

人工智能支架式学习中的公平性，通常被视为减少与既往经验、技能水平或教学支持相关的障碍 [37]。人工智能工具可以帮助学习者获取超出其当前语言熟练度的材料、弥补基础知识方面的不足，并克服残障相关障碍 [28]。关于人工智能与教育公平的研究，往往强调不同人口统计群体之间的可及性与公平性，包括性别、文化背景和社会经济地位 [52]。动机、认知负荷和情感投入等学习者状态，也可能影响可用性和学习成果 [10]。因此，要在实践中考察公平性，就需要关注学习者在使用过程中如何体验人工智能支持。

2.2 学习中的结构化视觉设计

人工智能支持的学习环境中的视觉支架已成为提升学习成果的有力策略，能够通过降低认知负荷并促进概念理解来实现这一目标。研究表明，交互式图表和模拟等视觉辅助工具，可以通过卸载认知需求，帮助学习者更有效地处理复杂信息 [36, 48]。结构化学习路径将抽象过程外化，以支持初学者。例如，通过将复杂算法分解为分步序列，学习者能够更加系统地获取知识 [21]。例如，基于步骤的动画通过展示算法决策的下一步可视化，辅助学生学习算法 [17]。然而，大多数关于视觉支架和结构化路径的研究都假设学习条件稳定且理想，学习者在这些条件下情绪投入、注意力集中并且能够进行工作

于最佳情境中。他们往往忽视学习者状态变化的影响，例如情绪脱离、分心或疲劳，而这些因素可能显著影响学习成果。先前研究表明，数字环境中的负面情绪会对学习成果产生不利影响 [65]。此外，使用过时设备或面临不可靠的互联网接入等欠佳学习环境，也可能损害学习有效性[51]。虽然结构化视觉设计通过认知可访问性促进包容性，但当引入情感和情境变化时，也可能带来隐藏成本。

2.3 情绪、可用性与交互脆弱性

学习者在深度学习活动中会经历动态的情绪波动 [10]。情绪具有首要的教育重要性，情感舒适度在塑造教育情境中的认知负荷、注意力和学习坚持性方面发挥着关键作用 [44]。研究表明，处于情感支持性环境中的学习者表现出更好的学习表现，因为积极情绪能够促进更深入的认知加工和动机 [26]。相反，焦虑、绝望、愤怒和无聊等负面情绪可能导致学习者失去学习兴趣。在情感计算领域，研究显示，挫败感、困惑或 disengagement 等负面情绪会显著降低感知学习系统可用性 [1]。例如，对数字学习系统的研究表明，困惑可能引发挫败感，并随着时间推移导致无聊 [1], ，最终降低用户对系统的信任以及对系统有效性的感知。然而，大多数教育人工智能系统并未明确考虑交互的情绪维度。尽管一些情感感知系统会采用情感计算技术，例如面部表情分析或生理传感器来检测情绪 [34], ，但大多数系统仍依赖标准化教学设计，并假设情绪状态保持稳定。这些系统通常根据系统性能、学习者成果和先验知识来计算教育人工智能系统的评估得分 [39], ，却没有考虑挫败感或 disengagement 等情绪状态可能如何影响这些指标。结构化辅导系统较为脆弱，因为当学习者偏离预设路径时，这些系统可能无法适应学习者的学习状态。例如， [17] 中提出的教育系统呈现固定的教学序列，通过交互式动画引导学习者完成每一个算法步骤。然而，任何环节的误解都可能妨碍对后续步骤的理解，而系统也缺乏复习或重复早期阶段的功能。因此，用户可能无法完成练习题等后续任务，进而导致交互中断，并使整体用户体验下降。

2.4 语境稳健性与设计中隐藏的假设

当代学习技术的设计与评估通常建立在对学习者情境的理想化假设之上。这些系统往往预设一种“数字富裕”环境——其特征是拥有可靠硬件、高速互联网接入以及充足的屏幕空间 [63]。然而，这类假设掩盖了教育技术的社会—技术纠缠，即技术使用嵌入于学习者混乱且动态的日常生活之中 [43]。因此，当针对大屏幕优化且依赖复杂视觉隐喻的界面被迁移到移动设备时，失效的不仅是布局，还有其底层的认知交互模型 [16]。更令人担忧的是，计算机教育研究中的效度评估长期以来偏向无菌实验室环境，造成了方法论情境盲点。这种评估范式无法捕捉非理想条件下的真实用户交互，例如情绪困扰、认知疲劳或环境干扰，而这些正是现实学习的典型特征 [56]。因此，那些在经过净化的环境中表现良好、在理论上看似“公平”的系统，在实际部署中可能会迅速退化，非但未能缓解教育不平等，反而会无意中加剧其

缓解作用 [54]。这些现象提醒我们，关于教育公平的讨论必须超越单纯的接入问题，转而关注技术在多样化现实情境中的“实践稳健性”。

2.5 迈向情绪韧性教育人工智能

为回应这些研究空白，本研究将评估重点从普适设计转向情境适应性，考察结构化、可视化的基于人工智能的教育工具在不同学习者背景和情感状态下的表现一致性。我们认为，尽管许多人工智能驱动的教育工具旨在通过提供高度结构化的支架式支持、降低认知进入障碍来促进公平，但这一策略掩盖了一种尚未得到充分审视的风险。这种做法可能潜藏着一种被忽视的风险，我们称之为公平—脆弱性张力：追求初始可用性和可及性（公平），可能会损害系统在复杂或有压力情境下的韧性与灵活性（脆弱性）。当学习者感到挫败或焦虑时，其认知资源会被耗尽 [44], ，原本旨在提供帮助的僵化支架可能转而成为认知约束，限制适应性反应，并削弱自我调节能力的发展 [27]。本研究的贡献在于为以公平为导向的设计引入一种更具动态性且具有生态效度的评价视角。我们认为，真正的公平不应局限于静态支持结构，而应涵盖具备适应性稳定性的系统——即能够敏锐回应学习者认知与情绪状态的变化，并在多变的现实条件下持续提供有效支持。因此，我们主张教育设计应从静态可用性转向动态韧性。

3 方法论

3.1 研究程序

本实验是西交利物浦大学（XJTLU）“高级面向对象编程”课程的一部分，共有320名具有编程经验的三年级学生参与。这一背景使他们能够学习相对复杂的算法概念。实验内容聚焦于三个图算法主题：图连通性、识别无向图中的连通分量，以及检测有向图中的环。这些主题是难度相当的典型图问题，涵盖有向图和无向图。它们需要运用计算思维并理解抽象的算法原理，因此适合用于评估本研究所考察的教学干预措施。本研究已获得XJTLU的伦理审批。

VCC 是 VisualCodeMOOC 的对话代理组件 [31], ，在结构化视觉人工智能条件下充当教学工具。它通过预先设定的一系列概念讲解、配有动态可视化的分步演示、代码讲解和练习题，引导学生完成图算法学习任务。可视化内容与聊天机器人的解释同步呈现算法步骤。在交互过程中，学生可以确认自己的理解、提出后续问题，并获得澄清说明以及对答案的反馈。为单独考察结构化视觉交互设计的影响，两种条件均通过相同的 ChatGPT 3.5 API 和模型配置，由同一大语言模型提供支持。两种条件在学习时长、教学主题及其顺序、学习内容与任务，以及教师指导的数量和时间安排方面均保持一致。在两种条件下，参与者都先针对第一个主题接受完整演示，再针对第二个主题接受部分指导，最后自主学习第三个主题。唯一有意设置的差异是交互与呈现形式：VCC 将学习过程组织为预先设定的阶段，并协调

聊天机器人解释与动态可视化；而标准人工智能条件则通过传统的基于文本的对话界面提供相应的学习支持。因此，两种条件之间的差异被解释为与交互和呈现形式有关，而非由底层语言模型、教学内容、接触时长或教师支持所致。

参与者被随机分配到两个组之一：VCC 组或标准 AI 组。为确保实验控制并尽量减少组间影响，参与者在不同时间进行的八场实验中完成研究。每场实验开始时，研究人员都会提供标准化说明，介绍学习任务和实验流程。每场实验约持续 50 分钟，并按照相同的学习主题顺序进行：（1）图连通性；（2）识别连通分量；（3）检测环。完成所有学习任务后，参与者填写研究后问卷。

对于标准 AI 组，教师辅助指导包括简要讲解算法概念，以及逐步指导如何使用 AI 工具进行学习。在第一个主题中，研究人员完整演示了如何使用 AI 工具学习算法。对于第二个主题，学生先接受部分指导，然后独立探索算法其余步骤的学习内容。在第三个主题中，学生使用 AI 工具进行自主学习。同样，对于 VCC 组，第一个主题包括全面演示系统界面和交互式学习过程。在第二个主题中，教师引导学生完成前两个阶段——理解算法概念及其基本步骤——之后，学生独立探索算法的其余步骤。第三个主题则作为一项完全自主学习任务进行。共收集到 320 份问卷答复。经过数据清理后，因存在可能无效作答的迹象（例如问卷各题答案模式一致或完成时间异常短），排除了 38 份答复。最终分析样本包含 282 份答复。在拥有完整人口统计数据的 282 名参与者中，83 人为女性，203 人为男性。其年龄范围为 18 至 26 岁（ $M = 21.09, SD = 0.91$ ）。

3.2 测量工具与适配

本研究采用随机化组间仅后测设计，比较两种人工智能支持的学习条件：VCC 和标准文本式 AI 聊天机器人。分配的 AI 条件构成自变量。主要因变量是参与者使用后的可用性、感知有效性、效率、满意度和情绪评分，使用五点李克特量表中的 50 个题项进行测量。研究收集了人口统计信息和学习者背景信息，以描述样本；同时，将先前的聊天机器人使用频率、先前与聊天机器人的情感体验以及设备类型作为子组变量进行考察。先前情感体验是指参与者此前对聊天机器人使用的感知，而情绪量表则评估参与者在学习环节结束后对所分配系统产生的情感反应。五个结果维度均代表参与者自我报告的感知，而非客观的学习表现。

我们设计的这份问卷建立在一个全面的多维框架之上，旨在系统评估人工智能聊天机器人在特定教育情境中的整体表现。其核心结构借鉴了经典的可用性标准 ISO 9241-11 [60], ，围绕有效性、效率和满意度这三个基本维度构建评估体系。在此基础上，问卷深入融合了技术接受模型（TAM）的核心理念 [7]，通过测量感知有用性（体现为有效性）和感知易用性（体现为可用性和效率），预测用户持续使用意愿。此外，为实现更全面的评估，该设计超越了传统的功能考量，融入了现代用户体验（UX）视角。借鉴用户体验问卷（UEQ）等模型， [30], 它

建立了专门的情感维度。该维度考察工具对学习者特征（如兴趣、动机和信心）的影响，同时评估其缓解挫败感的能力。

总体而言，这份问卷共包含 63 个条目。问卷首先收集参与者的人口统计信息及其背景资料，包括编程经验和以往使用人工智能聊天机器人的经验、学习风格、使用情境（学习地点和设备），以及使用人工智能工具的动机。问卷的核心部分由五个部分组成：可用性、有效性、效率、满意度和情绪，每部分包含 10 个问题，采用 5 点李克特量表评分（1 = 非常不同意； 5 = 非常同意）。这些部分分别用于评估工具的操作便捷性、其内容对学习的助益、系统响应表现及简化学习过程的能力、用户的总体满意度和继续使用意愿，以及对学习者动机和信心的积极情感影响。

3.3 问卷的效度与信度

为评估修订后问卷的构念效度，首先通过 SPSS 进行了探索性因子分析（EFA）。分析首先采用主成分分析（PCA），遵循 Kaiser 的建议 [22]，保留特征值大于 1.00 的因子。此外，还采用了 Cattell 碎石图准则 [5]。根据特征值和碎石图确定适当的提取因子数量。随后，进行了方差最大旋转的 PCA，以进一步探索潜在的因子结构。根据 DeCoster 提出的指导原则 [8],，采用以下标准保留因子：（1）每个条目的载荷应大于 0.4；（2）条目只能载荷于一个因子；（3）保留的因子与条目在概念上应保持一致；（4）每个因子至少应包含三个条目。

为评估提取因子的内部一致性信度，计算了 Cronbach α系数（α）。按照广泛认可的 0.70 [41]阈值，本研究采用 0.70 作为可接受信度的最低标准。可用性（α = .842，6 个条目）、有效性（α = .823，7 个条目）、效率（α = .815，7 个条目）、满意度（α = .935，8 个条目）和情绪（α = .876，5 个条目）均超过了可接受阈值。

为进一步检验问卷各构念的因子效度，本研究使用 AMOS，通过结构方程模型（SEM）进行验证性因子分析（CFA）。根据以下常用指标评估模型拟合度：(1) 卡方与自由度之比应满足 $\chi^2/df < 5$；(2) 近似误差均方根应满足 RMSEA < 0.08；(3) 标准化均方根残差应满足 SRMR < 0.08；(4) Tucker–Lewis 指数和比较拟合指数应分别满足 TLI > 0.90和 CFI > 0.90。验证性因子分析结果表明，模型拟合度可接受： $\chi^2 = 942.113$, $df = 485$, $\chi^2/df = 1.943$, SRMR $= 0.0546$, RMSEA $= 0.058$, TLI $= 0.901$, 以及 CFI $= 0.909$。相应的结构方程模型见图 1。

3.4 数据分析

使用 SPSS 和 AMOS 对问卷数据进行分析。完成数据清理后，根据保留条目计算五个评价维度的量表得分，得分越高表示评价越积极。通过探索性因子分析、验证性因子分析和 Cronbach α系数检验测量质量。采用皮尔逊相关考察五个维度之间的关联。进行单因素多元方差分析（MANOVA），检验 AI条件对可用性、有效性、效率、满意度和情绪的总体影响。随后采用独立样本 t-检验，识别各个维度上的差异；在适用情况下，报告偏 η^2 和 Cohen's d 作为效应量。

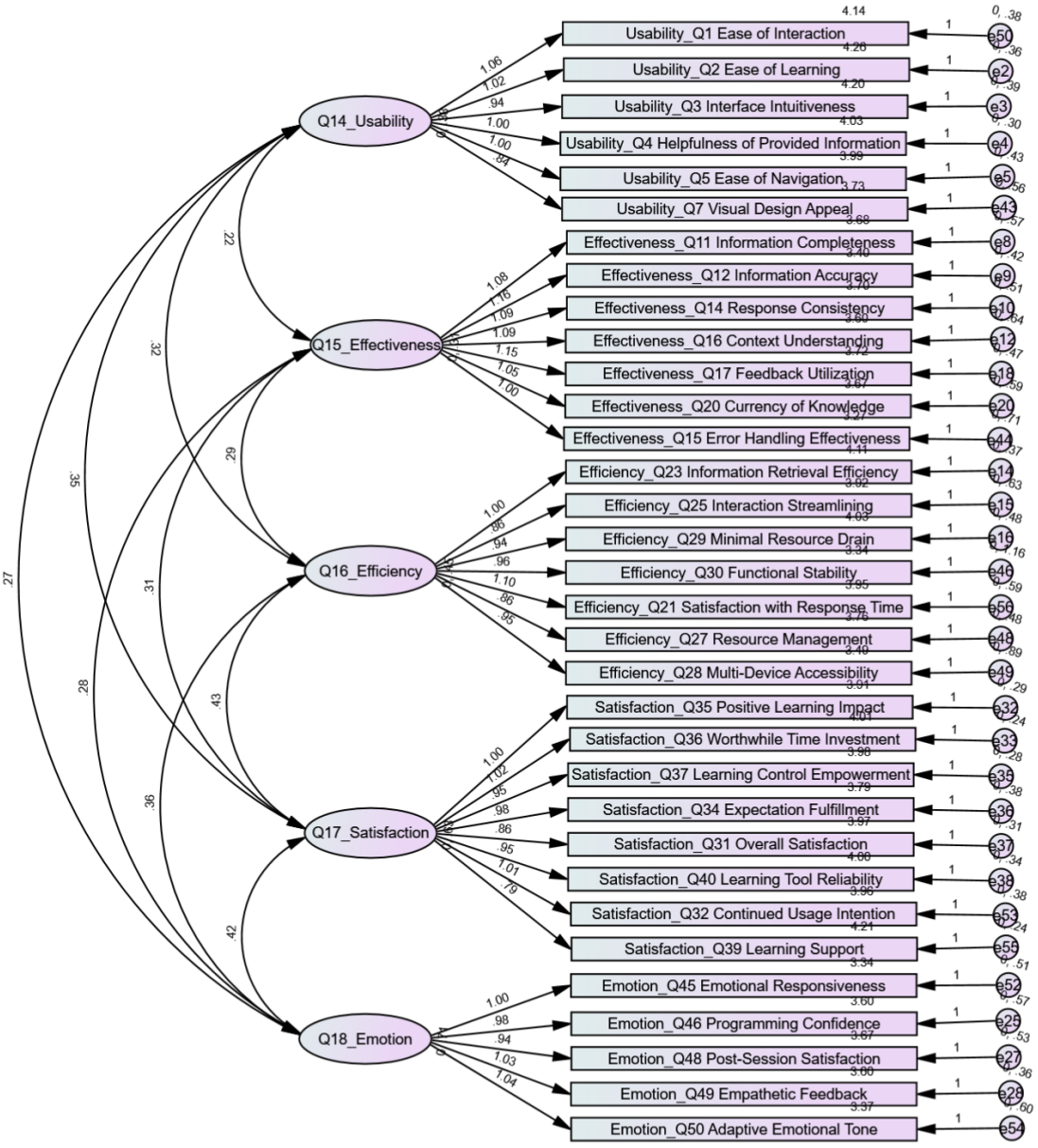

图1. 五个用户体验量表的验证性因子分析

为检验各组在聊天机器人评估维度上的差异，采用了两步分析方法。首先，在每种AI条件下分别进行多变量方差分析（MANOVAs），以同时评估各子组变量在多个因变量上的总体效应。对于未产生统计显著多变量效应的子组变量，不再进行后续分析，以聚焦于具有实际意义的差异。对于MANOVA结果显著的子组变量，则针对每个相关维度进行后续单变量方差分析（ANOVAs）。随后，采用经邦费罗尼校正的事后比较，以确定各组在因变量上的差异。当莱文检验表明违反方差齐性假设（$p < .05$）时，在调整后的假设下重新进行相应的ANOVA，并使用盖姆斯–豪厄尔事后检验，以适应方差和样本量不相等的情况。报告了Cohen's d 作为成对比较的效应量。我们采用 $\alpha = .05$ 的显著性阈值和双侧 t检验。

4 研究结果

我们首先考察了两种 AI条件在五个用户体验量表上的差异，随后分析了各量表之间的相关性。接着，我们在各条件内进行分析，以考察参与者的评价是否因先前的聊天机器人使用频率、先前与聊天机器人的情感体验以及设备类型而有所不同。所有显著的亚组比较结果均汇总于表 1。

4.1 StandardAI 与 VCC 之间用户体验的总体差异

首先，我们进行了一元多元方差分析，以考察 AI条件在五个用户体验维度上的联合效应。图 2 展示了被试间效应的单变量检验结果。基于未经调整的 p 值，检验结果表明，在名义 .05 阈值下，可用性存在差异， $F(1, 280) = 9.81$， $p = .002$，偏 $\eta^2 = .034$，有效性也存在差异， $F(1, 280) = 5.38$， $p = .021$，偏 $\eta^2 = .019$。分配至 VCC 条件的参与者报告的可用性和有效性高于分配至标准人工智能条件的参与者。

在有效性、 $F(1, 280) = 2.61, p = .108$，偏$\eta^2 \eta^2 = .009$；满意度、 $F(1, 280) = 2.98, p = .085$，偏$\eta^2$ $\eta^2 = .011$；或使用后情绪量表、 $F(1, 280) = 0.33, p = .565$，偏$\eta^2$ $\eta^2 = .001$方面，均未发现统计显著的被试间效应。从描述性统计来看，在全部五个量表中，标准人工智能条件下评分的标准差均小于VCC条件。

后续的独立样本 t-检验（图3）再次呈现了这一模式。由于均值差异的计算方式为标准人工智能减去VCC，因此负值表示VCC的评分更高。在可用性方面，差异具有统计显著性， $t(280) = -3.132, p = .002$，均值差异为 $= -0.24050, SE = 0.07680$，95%置信区间为 $[-0.39167, -0.08933]$；在有效性方面，差异也具有统计显著性， $t(280) = -2.319, p = .021$，均值差异为 $= -0.18348, SE = 0.07911$，95%置信区间为 $[-0.33921, -0.02775]$。在效率方面，差异未达到统计显著， $t(280) = -1.614, p = .108$，均值差异为 $= -0.13609$，95%置信区间为 $[-0.30203, 0.02985]$；满意度方面， $t(280) = -1.728, p = .085$，均值差异为 $= -0.15894$，95%置信区间为 $[-0.34006, 0.02217]$；情绪方面， $t(280) = -0.577, p = .565$，均值差异为 $= -0.06205$，95%置信区间为 $[-0.27390, 0.14981]$。五个量表的莱文检验均不显著（$p = .159$–$.784$），支持本文报告的等方差估计。

4.2 用户体验维度之间的关系

五个量表均呈正相关，且所有两两关联均达到统计显著性 $p < .001$（图4）。可用性与有效性（$r = .548$）、效率（$r = .634$）、满意度（$r = .642$）和情绪（$r = .454$）相关。有效性与效率（$r = .633$）、满意度（$r = .621$）和情绪（$r = .503$）相关。

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	Usability	4.073[a]	1	4.073	9.807	.002	.034
	Effectiveness	2.370[b]	1	2.370	5.379	.021	.019
	Efficiency	1.304[c]	1	1.304	2.606	.108	.009
	Satisfaction	1.779[d]	1	1.779	2.984	.085	.011
	Emotion	.271[e]	1	.271	.332	.565	.001
Intercept	Usability	4640.185	1	4640.185	11174.254	<.001	.976
	Effectiveness	3618.530	1	3618.530	8211.117	<.001	.967
	Efficiency	4087.628	1	4087.628	8169.291	<.001	.967
	Satisfaction	4470.087	1	4470.087	7499.695	<.001	.964
	Emotion	3499.900	1	3499.900	4291.446	<.001	.939
Subgroup	Usability	4.073	1	4.073	9.807	.002	.034
	Effectiveness	2.370	1	2.370	5.379	.021	.019
	Efficiency	1.304	1	1.304	2.606	.108	.009
	Satisfaction	1.779	1	1.779	2.984	.085	.011
	Emotion	.271	1	.271	.332	.565	.001
Error	Usability	116.272	280	.415			
	Effectiveness	123.392	280	.441			
	Efficiency	140.102	280	.500			
	Satisfaction	166.890	280	.596			
	Emotion	228.355	280	.816			
Total	Usability	4776.139	282				
	Effectiveness	3755.429	282				
	Efficiency	4239.367	282				
	Satisfaction	4650.719	282				
	Emotion	3735.120	282				
Corrected Total	Usability	120.344	281				
	Effectiveness	125.763	281				
	Efficiency	141.406	281				
	Satisfaction	168.669	281				
	Emotion	228.626	281				

a. R Squared = .034 (Adjusted R Squared = .030)

b. R Squared = .019 (Adjusted R Squared = .015)

c. R Squared = .009 (Adjusted R Squared = .006)

d. R Squared = .011 (Adjusted R Squared = .007)

e. R Squared = .001 (Adjusted R Squared = -.002)

图2. 五个用户体验维度的被试间效应检验

效率与满意度的观测关联最强（r = .696），且也与情绪相关（r = .555）。满意度与情绪呈正相关（r = .584）。

对于亚组分析，在每种AI条件下，分别针对先前聊天机器人使用频率、先前与聊天机器人的情感体验以及设备类型进行了多变量方差分析。显著的多变量效应随后通过单变量方差分析和校正后的两两比较进行检验。总体检验、描述性统计、假设检验和事后程序汇总于表1中；下文重点介绍显著的两两比较结果。

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Usability	Equal variances assumed	.856	.356	-3.132	280	.002	-.24050	.07680	-.39167	-.08933
	Equal variances not assumed			-3.120	271.558	.002	-.24050	.07708	-.39226	-.08874
Effectiveness	Equal variances assumed	.476	.491	-2.319	280	.021	-.18348	.07911	-.33921	-.02775
	Equal variances not assumed			-2.329	279.317	.021	-.18348	.07877	-.33855	-.02841
Efficiency	Equal variances assumed	1.992	.159	-1.614	280	.108	-.13609	.08430	-.30203	.02985
	Equal variances not assumed			-1.610	273.900	.109	-.13609	.08453	-.30251	.03033
Satisfaction	Equal variances assumed	.076	.784	-1.728	280	.085	-.15894	.09201	-.34006	.02217
	Equal variances not assumed			-1.729	279.149	.085	-.15894	.09195	-.33995	.02207
Emotion	Equal variances assumed	1.296	.256	-.577	280	.565	-.06205	.10762	-.27390	.14981
	Equal variances not assumed			-.579	279.570	.563	-.06205	.10720	-.27307	.14898

图3. 标准人工智能与VCC的独立样本检验

Correlations

		Usability	Effectiveness	Efficiency	Satisfaction	Emotion
Usability	Pearson Correlation	--				
	N	282				
Effectiveness	Pearson Correlation	.548**	--			
	Sig. (2-tailed)	<.001				
	N	282	282			
Efficiency	Pearson Correlation	.634**	.633**	--		
	Sig. (2-tailed)	<.001	<.001			
	N	282	282	282		
Satisfaction	Pearson Correlation	.642**	.621**	.696**	--	
	Sig. (2-tailed)	<.001	<.001	<.001		
	N	282	282	282	282	
Emotion	Pearson Correlation	.454**	.503**	.555**	.584**	--
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	
	N	282	282	282	282	282

**. Correlation is significant at the 0.01 level (2-tailed).

图4. 五个用户体验维度之间的皮尔逊相关

表1. 按维度和AI条件划分的方差分析汇总

维度	**问题**	**组1（标准人工智能）**	**组2（VCC）**

可用性	**交互**	**过去对聊天机器人的感受** 感受中立的参与者报告称 交互显著更低 评分低于感到舒适的参与者。	**过去对聊天机器人的感受** 感到不舒适的参与者报告称 交互评分在统计上显著更低 评分低于感到中立的参与者。
	用户采纳	**以往聊天机器人使用经验** 很少使用聊天机器人的参与者 报告称，其用户 采用度评分显著低于那些偶尔 使用、经常使用和每天使用聊天机器人的参与者。 **过去对聊天机器人的感受** 感到中立的参与者报告称 其用户采用度评分在统计上显著更低 于感到舒适的参与者。	Null
	界面	Null	**聊天机器人设备** 使用笔记本电脑的参与者报告称 统计显著更高的界面 评分，相较于使用平板电脑或 智能手机的参与者。 **过去对聊天机器人的感受** 感到不舒适的参与者报告称 统计显著更低的界面 评分低于那些感到舒适的人。
	布局	**过去对聊天机器人的感受** 感到中立的参与者报告称 布局评分在统计上显著低于 那些感到舒适的人。	**过去对聊天机器人的感受** 感到不舒适的参与者报告称 布局评分在统计上显著更低 相比之下，感到中立和舒适的参与者评分更高。
有效性信息	**完整性**	**过去对聊天机器人的感受** 感到中立的参与者报告称 信息完整性评分在统计上显著低于 那些感到 舒适。	**过去对聊天机器人的感受** 感到不舒适的参与者报告称 其信息 完整性评分显著低于感到 中立。
	响应 一致性	**过去对聊天机器人的感受** 感受中立的参与者报告称， 其响应一致性评分在统计上显著低于 感受舒适的参与者。 舒适。	Null
效率	**响应时间**	**过去对聊天机器人的感受** 感到中立的参与者报告称 响应时间得分在统计上显著较低 相比之下，感到舒适的参与者得分更高。	**过去对聊天机器人的感受** 感到不舒适的参与者报告称 响应时间得分在统计上显著更低 ，低于感到中立和 舒适。

	信息检索	**过去对聊天机器人的感受** 感到不舒适和 中立地报告了具有统计显著性的结果 信息检索得分低于那些 感到舒适的人。	Null
满意度	**学习管理**	**以往聊天机器人使用经验** 很少使用聊天机器人的参与者 报告的得分统计显著较低 学习管理得分低于那些 经常使用以及每天使用聊天机器人的参与者。 **过去对聊天机器人的感受** 感受中立的参与者报告称 统计显著较低的学习 管理得分，而那些感到 舒适。	**过去对聊天机器人的感受** 感到不舒适的参与者报告称 统计显著较低的学习 管理得分低于那些感到 舒适的人。

4.3 先前聊天机器人使用频率与用户采纳和学习管理评价

在标准人工智能条件下，先前的聊天机器人使用频率与两项评价相关。很少使用聊天机器人的参与者，其用户采用度评分低于偶尔使用的参与者（均值差异 $= -1.091$, $p < .05$，Cohen's $d = 0.7603$）、经常使用的参与者（均值差异 $= -1.117$, $p < .01$, $d = 0.7603$）或每天使用的参与者（均值差异 $= -1.486$, $p < .001$, $d = 0.7603$）。很少使用聊天机器人的参与者，其学习管理得分也低于经常使用的参与者（均值差异 $= -0.942$, $p < .05$, $d = 0.7092$）或每天使用的参与者（均值差异 $= -1.382$, $p < .01$, $d = 0.7092$）。未报告与 VCC 相关的频率差异比较达到显著水平。

4.4 先前情感体验与跨用户体验维度的评价

参与者的先前聊天机器人使用情感体验与对两个系统的评价均相关，但显著成对差异所在的位置因条件而异。在标准人工智能条件下，显著差异通常出现在报告中立和舒适的先前体验的参与者之间。在 VCC 条件下，显著差异主要集中在报告不舒适的先前体验的参与者中。

*4.4.1 交互与界面评价。*对于标准人工智能条件，具有中立先前聊天机器人使用经历的参与者，其交互评分低于具有舒适经历的参与者（平均差异 $= -0.514$、$p < .05$、$d = 0.5211$）。与具有舒适先前体验的参与者相比，他们的用户采纳评分（平均差异 $= -0.646$、$p < .01$、$d = 0.6074$）和布局评分（平均差异 $= -0.573$、$p < .05$、$d = 0.4885$）也更低。对于界面，未报告与情感相关的显著比较结果。

对于 VCC，与具有中性的既往聊天机器人使用经历的参与者相比，具有不舒适的既往聊天机器人使用经历的参与者报告的交互评分更低（平均差异 $= -1.099$、$p < .01$、$d = 0.5800$）。与具有舒适经历的参与者相比，同一组参与者对界面的评分也更低（平均差异 $= -0.703$、$p < .05$、$d = 0.4304$）。与具有中性经历（平均差异 $= -0.941$、$p < .05$、$d = 0.4843$）或舒适经历（平均差异 $= -0.827$、$p < .05$、$d = 0.4843$）的参与者相比，他们报告的布局评分也更低。对于 VCC，未报告与情感相关的用户采纳比较结果。

4.4.2 信息质量与效率评估。 在标准人工智能条件下，与具有舒适经历的参与者相比，具有中性的既往聊天机器人使用经历的参与者对信息完整性的评分更低（平均差异= −0.882、 $p < .001$、 $d = 0.8189$）。他们对响应一致性（平均差异 $= -0.577$、$p < .05$、 $d = 0.5528$）和响应时间（平均差异 $= -0.813$、 $p < .01$、 $d = 0.6385$）的评分也更低。对于信息检索，与具有舒适经历的参与者相比，具有不舒适经历（平均差异 $= -1.200$、$p < .01$、 $d = 1.0263$）或中性既往经历（平均差异 $= -1.000$、 $p < .001$、 $d = 1.0263$）的参与者报告的评分更低。

对于VCC，具有不舒适的先前体验的参与者报告的信息完整性低于具有中立体验的参与者（均值差异 $= -1.033$、 $p < .05$、 $d = 0.4487$）。与具有中立体验（均值差异 $= -1.013$、 $p < .05$、 $d = 0.4665$）或舒适体验（均值差异 $= -0.960$、 $p < .05$、 $d = 0.4665$）的参与者相比，他们对响应时间的评价也更低。在VCC条件下，未报告响应一致性或信息检索方面具有显著性的情感相关比较。

4.4.3 学习管理满意度。在标准人工智能条件下，具有中立既往聊天机器人使用经历的参与者报告的学习管理得分低于具有舒适的先前体验的参与者（均值差异 $= -1.017$、$p < .001$、 $d = 0.9012$）。在VCC条件下，具有不舒适的先前体验的参与者报告的学习管理得分低于具有舒适的先前体验的参与者（均值差异 $= -0.810$、$p < .05$、 $d = 0.4669$）。

4.5 设备类型与界面评估

在 VCC 条件下，设备类型与界面评分有关。使用笔记本电脑的参与者对界面的评价高于使用平板电脑或智能手机的参与者（平均差异 $= 0.770$, $p < .05$, $d = 0.4304$）。标准人工智能条件下未报告显著的设备相关比较。

5 讨论

本研究比较了标准人工智能与 VCC 在可用性、感知有效性、效率、满意度和使用后情感体验方面的表现。研究还考察了系统评价如何随先前的聊天机器人使用频率、先前与聊天机器人互动的情感体验以及设备类型而变化。本讨论探讨三个问题：结构化支架能提供哪些支持、其益处又有哪些局限；这些益处如何因学习者背景而异；以及使用条件如何制约包容性设计。

5.1 支架式支持的益处与局限

VCC 在可用性和感知有效性方面的评分显著高于标准人工智能，而在效率、满意度或使用后情感体验方面，整体差异均未达到显著水平。因此，结构化视觉支持的益处集中体现在特定体验维度上，而非用户体验各方面的普遍提升。

可用性优势可能反映了 VCC 结构化的教学序列以及高度整合的可视化。可预测的学习步骤有助于明确后续操作，而分步可视化则提供情境化反馈，使抽象内容（例如算法过程）变得更加具体且可观察。这一解释与 Naps 等人关于算法可视化教育价值的论述相呼应，即算法可视化能够积极调动学习者 [40]。相比之下，标准人工智能更加开放式的对话和文本解释可能要求学习者自行提出查询，并跟踪算法状态的变化。Sweller 关于教学结构和认知负荷的研究提供了互补性解释：引导不足或结构不佳的材料可能增加外在认知

负荷 [57]。因此，VCC 的可用性优势可能反映了将部分组织学习和跟踪状态的工作转移给界面的价值，尽管这一认知机制并未被直接测量。

感知有效性方面的优势同样可以通过外部表征和心智模型构建来理解。可视化能够将抽象描述与计算过程的动态行为联系起来，从而支持学习者对算法的概念理解 [21]。将学习过程分段并外化状态变化，可能会降低追踪瞬时信息时的工作记忆负荷，使更多认知资源可用于理解底层逻辑。这一解释与认知负荷理论一致 [6]。Moreno 关于交互式多媒体学习的研究也支持循序渐进式参与可能有助于理解和心智模型构建这一观点 [38]。不过，由于有效性是通过学习者的自我报告进行评估的，这些发现表明的是学习者感知到的更大支持，而非在知识获取、迁移或长期学习方面得到证实的收益。

结构化支架可能也会引入交互约束。效率差异不显著的一种可能解释是，循序渐进的序列能够提供清晰的学习路径，但未必能满足快速查询或直接访问特定难点的需求。Brusilovsky 关于自适应超媒体的研究表明，引导式路径能够支持新手，而经验丰富的用户或具有特定目标的用户可能更加重视非线性访问和控制 [4]。学习解读可视化内容也需要时间和认知努力 [58]。这些要求可能限制可用性优势转化为感知效率的程度，尽管本研究并未直接检验用户控制或学习视觉表征所需努力的作用。

满意度方面的研究结果进一步表明，清晰度只是评价交互的一个依据。正如关于用户控制和满意度的研究所指出的，受限的学习序列、未满足的交互预期或控制感的减弱，都可能影响用户对结构化系统的整体评价 [42]。标准人工智能的熟悉性和灵活性也可能提升其情感体验价值。在Hassenzahl的论述中，用户体验同时包含工具性质量和非工具性质量；因此，可用性的提升未必会带来满意度的相应提高 [18]。尽管这一不显著结果并不能证明两种工具的满意度相当，但它凸显了同时考虑支架式支持所提供的指导以及学习者对自主控制的需求的重要性。

使用后情感体验不存在显著差异，同样表明，改变信息呈现方式可能不足以产生可检测的情感收益。算法学习可能涉及困难、困惑和挫败感，而Pekrun等人关于学业情绪的研究强调了认知评估和动机因素在情感体验中的作用 [45]。本研究的比较并未证明任一系统都无法改善学习者的情绪；它仅表明未观察到显著差异。关于结对编程等协作式方法与编程教育中焦虑和挫败感之间关系的研究[53], 表明，情感支持还可能涉及社会交互，以及超越信息呈现的学习组织。这些可能性值得进一步研究。

五个体验维度之间的正相关提供了一个互补视角。可用性与感知有效性、效率和满意度密切相关，其中效率与满意度之间的关联最强。因此，学习者的评价包含多个相互关联的维度。感知有效性与效率之间的关联，可以结合 Finneran 关于目标达成和交互体验的讨论来考量 [12], ，而 Tractinsky 关于美学和可用性判断的研究则提醒我们，不应将整体体验简化为单一指标 [59]。这些关联也与技术接受模型中的易用性和有用性之间的关系相呼应 [61]。然而，相关性并不能证明

改善一个维度会带动其他维度相继改善，也无法解释系统间差异背后的因果机制。综合来看，研究结果表明，结构化支架有助于形成更清晰的交互，使学习者认为其更有帮助；但它对效率、满意度和情感的益处仍需进一步考察。

5.2 经验与情感脆弱性

除了整体均值之外，亚组研究结果表明，应结合学习者的既往经验来解读结构化支架的益处。在标准人工智能条件下，很少使用聊天机器人的参与者对用户采纳和学习管理的评分较低，而VCC在使用频率方面未显示出显著差异。这一模式与以下可能性相符：结构化支持降低了与经验相关的障碍；但这并不能证明VCC消除了与既往经验相关的差异。

开放式聊天机器人交互通常要求用户提出问题、引导对话，并评估回复是否满足其学习需求。因此，既往使用经验可能与更顺畅的采纳和学习管理有关。界面体验和交互焦虑方面的研究为这一解读提供了背景[47]。通过预定义步骤、算法演示和同步可视化，VCC承担了部分组织学习的工作，并可能减轻独立构建交互路径的负担。这一解读与支架式支持在弥合知识鸿沟方面的教育作用相一致 [46]。尽管如此，本研究并未直接测量提示技能，也不应将使用频率等同于专业水平。先前研究还表明，对话代理的交互局限可能令经验丰富的用户也感到沮丧 [66], 因此，不应仅凭熟悉程度就认定用户会有良好的体验。

不同于使用频率所呈现的模式，先前情感体验与两个系统中的多个评价维度均存在关联。在此，情感背景指参与者先前使用聊天机器人的舒适或不舒适体验，这不同于本次交互结束后施测的情绪量表。在标准人工智能条件下，显著差异通常出现在报告中性的先前体验与舒适的先前体验的参与者之间。在 VCC 中，较低的评价则更多集中于报告不舒适的先前体验的参与者。这些模式表明，降低与经验相关的进入门槛，并不一定能确保具有不同情感背景的学习者获得可比的交互体验。

这一模式体现在交互和布局评价中。在标准人工智能条件下，具有中性的先前体验的参与者对交互和布局的评价低于具有舒适的先前体验的参与者。在 VCC 中，具有不舒适的先前体验的参与者对交互和布局的评价低于具有中性或舒适的先前体验的参与者，并且对界面的评价低于具有舒适的先前体验的参与者。关于情感和注意资源的研究 [14]„ 结合加工流畅性理论 [49], , 为解释这一现象提供了可能的视角：学习者既有的情感评估和交互预期，可能与他们理解并接受新界面的方式有关。不过，本研究并未测量注意力的瞬时变化、迷失方向感或情感反馈回路，因此，这些机制仍只是可能的解释，而非已经确立的研究发现。

信息完整性呈现出相关模式。在标准人工智能中，既往经验为中立的参与者评分低于经验舒适的参与者；在 VCC 中，既往经验不舒适的参与者对完整性的评分低于经验中立的参与者。关于情感与信息搜寻的研究表明，情感评估可能与用户如何判断信息是否满足其目标有关 [29]。结构化呈现也可能通过帮助用户查找和组织内容，提升其对信息完整性的感知 [35]。然而，

VCC 中与既往经验不舒适相关的剩余差异表明，清晰的内容组织并不意味着所有学习者都认为自己获得了充分的学习支持。

响应一致性和效率方面的发现进一步表明，这一模式并未延伸至每一项测量指标。在标准人工智能中，具有中性的先前体验的参与者对响应一致性的评价低于具有舒适体验的参与者，而在 VCC 中则未报告相应的显著比较。情感注入模型为情感背景与评价判断之间的关系提供了理论视角 [13],但当前研究结果并不支持将一致性差异归因于极端情绪或压力下的认知变化。VCC 预先设定的序列可能使交互更具可预测性，尽管缺乏显著差异并不能证明不同情感背景下的响应信度相同。可预测性与交互自然度之间的关系仍有待进一步研究 [33]。

在两个系统中，响应时间评价都与先前情感体验相关，尽管显著比较涉及不同的子群体。仅在标准人工智能中报告了信息检索方面的显著差异：具有不舒适或中性的先前体验的参与者，其评分低于具有舒适体验的参与者。VCC 的序贯呈现可能减轻组织查询的负担，这与认知负荷理论关于结构化指导如何支持新手的解释一致 [23]。与此同时，信息检索研究表明，任务复杂性也可能影响用户体验 [32]。因此，当前结果最好被理解为主观效率评价与情感背景之间的关联。它们并不能证明两个系统具有相同的客观响应时间、更高的检索表现，或任一系统对情感的依赖显著减弱。

学习管理研究结果将这一模式延伸至满意度。标准人工智能显示出与使用频率和先前情感体验均相关的评价差异。对于 VCC，仅报告了与先前情感体验相关的显著比较：报告不舒适的先前体验的参与者，其评分低于报告舒适的先前体验的参与者。结合学习交互和情感投入方面的研究来看， [9],一种可能的解释是，预先设定的序列可以承担部分组织学习的工作，但未必能消解学习者对人工智能交互既存的不适或负面预期。这一区分凸显了分别考察对初始采用的支持，以及长期管理学习的体验的价值。

总的来说，这些研究结果为所提出的公平—脆弱性张力提供了探索性支持：结构化设计可能降低与既往使用经验相关的交互障碍，但仍会受到与不舒适的先前体验相关的评价差异影响。这里的公平仅限于不同经验背景的参与者获得交互支持的机会；这并不意味着更广泛的教育公平已经实现。同样，情感脆弱性是一种解释框架，而不是关于结构化设计必然造成情感伤害的结论。尤其是，一个系统内部存在显著差异，而另一个系统内部的差异不显著，这两点本身并不能构成系统在影响方面存在差异的统计证据。要确立这一张力，需要直接检验系统条件与学习者背景之间的交互作用，同时采用纵向或实验设计，将先前情感体验与使用过程中的情绪变化区分开来。

5.3 包容性设计的情境限制

应当结合访问结构化视觉支持所使用的设备来理解其价值。在 VCC 中，使用笔记本电脑的参与者对界面的评价高于使用平板电脑或智能手机的参与者；对于标准人工智能，则未报告具有显著性的设备相关比较。这一发现表明，视觉学习

界面的体验可能取决于其呈现方式和交互条件。但这并不能证明设备相关效应仅存在于 VCC，也不能证明两个系统对设备类型的敏感性存在差异。

显示空间提供了一种可能的解释。算法可视化可能要求学习者同时参考程序步骤、状态变化和文本解释，而有限的显示空间可能增加查找信息以及在不同表征之间切换所需的精力。关于小屏幕与图形信息理解的研究为这一解释提供了背景 [20]。然而，本研究比较了不同设备类型，却没有单独操纵或测量屏幕尺寸的影响。输入方式、布局适配以及周围的使用环境也可能影响界面评价。因此，研究结果揭示的是体验中与设备相关的差异，而非屏幕尺寸的因果效应。

这一结果为前文关于体验和情感背景的讨论增添了一项情境边界。即便系统提供了清晰的支架式支持，或表现出较少与经验相关的差异，其支持仍需通过特定设备和交互条件来实现。因此，包容性设计不能仅通过平均可用性或初始采用的障碍来评估；还需要考察不同学习者能否在各种使用条件下持续获得支持。本研究分别发现了与情感背景和设备类型相关的评价差异，但没有检验二者的联合作用。因此，研究并未证实小屏幕会加剧情感脆弱性。

基于这一视角，我们主张将关注点从初始可及性拓展至持续支持，并使用“韧性”来描述系统在学习者状态和使用条件发生变化时维持支持的能力。韧性是由这些发现所促成的一种理论与评估方向，而不是本研究所测量的系统属性。未来的工作应结合跨设备比较、对持续使用的观察，以及交互过程中情感的测量，以考察结构化支架在不断变化的条件下能否持续发挥效益。因此，降低进入门槛和维持交互支持，逐渐成为教育人工智能相互关联的设计目标。下一节将进一步阐述它们对设计、教学和部署的影响。

6 条建议

VCC 在可用性和感知有效性方面获得更高评分，表明结构化视觉支持具有针对性益处；与此同时，在效率、满意度或使用后情绪方面，未发现总体上的显著差异。然而，这些平均值并不能说明具有不同既往经验或使用不同设备的学习者是否能够继续有效地使用该支持。因此，我们建议将清晰的入门流程、对支持的持续可及性，以及经独立展示的理解视为相互关联的设计与部署目标。

首先，结构化可视化应促使学习者进行推理。既有研究表明，算法可视化的学习价值很大程度上取决于学习者如何与之互动 [21, 40]。设计者可以在现有活动中嵌入简短的“预测—观察—解释”循环。例如，在环检测任务中，学习者可以预测接下来将访问哪个节点或边，观察由此产生的状态变化，并解释任何不一致之处。这样一来，可视化便成为检验推理的工具，而非取而代之的手段。由于本研究测量的是感知到的学习成果，而非客观学习成果，因此应将这种模式作为一种候选学习机制进行检验。

其次，当推理陷入困境时，学习者应拥有清晰且由学习者控制的恢复方式。系统不应强迫学习者按照固定序列继续推进，而可以允许学习者重新检查当前步骤、请求局部提示，或在保留进度的同时返回之前的图状态。这一方向与研究中涉及既往聊天机器人使用情况和使用聊天机器人的先前情感体验的组内关联相一致。既有研究同样强调，需要在获得辅助与持续认知投入之间取得平衡 [25, 27]。辅助应帮助

学习者解决有益的困惑，而不是消除所有困难 [11]。当前数据支持将学习者控制的辅助作为一种设计方向，但不支持自动情绪识别或基于情感的干预。

第三，工具支持的学习应与对概念理解的独立检查相结合。两种研究条件都包括教师示范、部分指导和后续自主使用，因此这些建议适用于有组织的课堂部署。教育工作者可以示范如何提出聚焦推理的问题，以及如何请求澄清，但不应将感知有效性视为概念掌握。完成一项人工智能支持的活动后，教师可以让学习者在不使用人工智能的情况下解决一个新的图问题，并解释深度优先搜索如何识别环路，以及相关状态为何会发生变化。相关研究表明，学习者体验、任务表现和独立理解应分别进行评估 [2, 3]。因此，部署决策应同时考虑学习者体验和独立表现。

第四，机构应评估支持在具有代表性的学习条件下是否易于使用，而不仅仅是是否提供了访问权限。课程级试点应纳入既往聊天机器人使用经历和设备各不相同的学习者，并记录他们是否能够完成关键步骤、请求帮助后是否能够继续学习、他们如何体验交互，以及之后能否在不使用人工智能的情况下展示自己的理解。这体现了教育技术中访问、使用和结果之间的区别 [63]。如果某些设备或布局屡屡妨碍核心任务，机构就应改进界面；与此同时，应提供设备或替代途径，以支持相同的学习目标。收集任何交互或情感相关数据，也应有明确的目的和保护计划。

最后，应将这些做法视为可检验的设计与部署假设。未来研究应直接检验系统条件与既往聊天机器人使用情况、先前情感体验和设备类型之间的交互作用，而不是根据某个子组显著、另一个子组不显著来推断差异 [15]。研究还应考察重复使用、中断后恢复、任务中体验，以及延迟评估或无人工智能评估。实践目标是在检验这些支持是否能促使学习者展现独立理解能力的同时，确保不同经历和设备条件下的学习者都能寻求并利用支持。

7 结论

在我们的研究中，VCC 在可用性和感知有效性方面获得了高于标准人工智能的评价，而在效率、满意度或使用后情绪方面未发现显著差异。在标准人工智能条件下，较少使用聊天机器人的用户对用户采纳和学习管理的评价较低；相比之下，VCC 未报告相应的显著频率相关比较。在 VCC 条件下，报告有不舒适的既往聊天机器人使用经历的学习者，对交互和布局的评价低于其他学习者；笔记本电脑用户对界面的评价则高于平板电脑或智能手机用户。总体而言，这些探索性模式区分了两个问题：一是能否降低与经历相关的进入障碍，二是支持是否能在不同情感背景和使用条件下保持可用。因此，所提出的公平—脆弱性张力是一种评价视角，而不是由这些数据确立的因果效应。未来的评价应直接检验条件与背景之间的交互作用，考察不同设备上的使用体验，并独立于自我报告评估学习效果。归根结底，对结构化支持的判断应取决于学习者能否在不同情境下可靠地使用它，以及能否在人工智能交互之外展示理解。

致谢

我们衷心感谢所有参与者为本研究付出的时间和作出的贡献。同时，我们也感谢评审者提出的深思熟虑的反馈和宝贵建议，这些意见改进了本项工作。

参考文献

[1] Amaël Arguel、Lori Lockyer、Ottmar V. Lipp、Jason M. Lodge 和 Gregor Kennedy。2017。《由内而外：检测学习者的困惑以改进交互式数字学习环境》。《教育计算研究杂志》55，4（2017），526–551。doi:10.1177/0735633116674732[2] Patrick Bassner、Ben Lenk-Ostendorf、Ramona Beinstingel、Tobias Wasner 和 Stephan Krusche。2026。《更少压力，更高分数，相同学习：人工智能支持的编程教育中表现与学习的分离》。《计算机与教育：人工智能》10（2026），100537。doi:10.1016/j.caeai.2025.100537[3] Hamsa Bastani、Osbert Bastani、Alp Sungu、Haosen Ge、Özge Kabakcı 和 Rei Mariman。2025。《缺乏防护机制的生成式人工智能可能损害学习：来自高中数学的证据》。《美国国家科学院院刊》122，26（2025年7月），e2422633122。doi:10.1073/pnas.2422633122[4] Peter Brusilovsky。2001。《自适应超媒体》。《用户建模与用户适应性交互》11，1-2（2001年3月），87–110。doi:10.1023/a:1011143116306[5] Raymond B. Cattell。1966。《用于确定因子数量的碎石检验》。《多元行为研究》1，2（1966），245–276。doi:10.1207/ s15327906mbr0102 10_[6] Paul Chandler 和 John Sweller。1991。《认知负荷理论与教学形式》。《认知与教学》8，4（1991年12月），293–332。doi:10.1207/s1532690xci0804 2_[7] Fred D. Davis。1989。《感知有用性、感知易用性与信息技术的用户接受度》。《管理信息系统季刊》13，3（1989），319–340。doi:10.2307/249008[8] Jamie DeCoster。1998。《因子分析概述》。技术报告。阿拉巴马大学心理学系。https://www.stat- help.com/factor.pdf[9] Sidney D’mello 和 Art Graesser。2012。《AutoTutor 与情感型 AutoTutor》。《ACM交互式智能系统汇刊》2，4（2012年12月），1–39。doi:10.1145/2395123.2395128[10] Sidney D’Mello 和 Art Graesser。2012。《复杂学习过程中的情感状态动态》。《学习与教学》22，2（2012），145–157。doi:10.1016/j.learninstruc.2011.10.001[11] Sidney D’Mello、Blair Lehman、Reinhard Pekrun 和 Art Graesser。2014。《困惑有益于学习》。《学习与教学》29（2014年2月），153–170。doi:10.1016/j.learninstruc.2012.05.003[12] Christina M. Finneran 和 Ping Zhang。2003。《计算机媒介环境中心流前因的人–物–任务（PAT）模型》。《国际人机交互研究杂志》59，4（2003年10月），475–496。doi:10.1016/s1071-5819(03)00112-5[13] Joseph P. Forgas。1995。《情绪与判断：情感注入模型（AIM）》。《心理学通报》117，1（1995），39–66。doi:10.1037/0033- 2909.117.1.39[14] Philip Gable 和 Eddie Harmon-Jones。2010。《忧郁拓展，而厌恶收窄》。《心理科学》21，2（2010年1月），211–215。doi:10.1177/0956797609359622[15] Andrew Gelman 和 Hal Stern。2006。《“统计显著”与“不显著”之间的差异本身并不具有统计显著性》。《美国统计学家》60，4（2006年11月），328–331。doi:10.1198/000313006X152649[16] Michail N. Giannakos 和 Panayiotis Vlamos。2013。《在教育中使用网络广播：对其有效性的评估》。《英国教育技术杂志》44，3（2013），432–441。doi:10.1111/j.1467-8535.2012.01309.x[17] Foteini Grivokostopoulou、Isidoros Perikos 和 Ioannis Hatzilygeroudis。2017。《用于学习搜索算法并自动评估学生表现的教育系统》。《国际人工智能教育杂志》27，1（2017），207–240。doi:10.1007/s40593- 016-0116-x[18] Marc Hassenzahl。2018。《物与我：理解用户与产品之间的关系》。载于《Funology 2》。施普林格国际出版公司，沙姆，瑞士，301–313。doi:10.1007/978-3-319-68213-6_19[19] Wayne Holmes 和 Ilkka Tuomi。2022。《教育中的人工智能：现状与实践》。《欧洲教育杂志》57，4（2022），542–570。doi:10.1111/ejed.12533[20] Yu-Chen Hsieh、Chien-Ting Kuo 和 Hsuan Lin。2016。《移动设备屏幕尺寸对阅读效率的影响》。载于《老年人群信息技术的人类因素：适老化设计》，Jia Zhou 和 Gavriel Salvendy（编）。《计算机科学讲义》，第9754卷。施普林格，沙姆，瑞士，435–445。doi:10.1007/978-3-319-39943-0_42[21] CHRISTOPHER D. HUNDHAUSEN、SARAH A. DOUGLAS 和 JOHN T. STASKO。2002。《算法可视化有效性的元研究》。《可视语言与计算杂志》13，3（2002年6月），259–290。doi:10.1006/jvlc.2002.0237[22] Henry F. Kaiser。1960。《电子计算机在因子分析中的应用》。《教育与心理测量》20，1（1960），141–151。doi:10.1177/001316446002000116[23] Slava Kalyuga、Paul Chandler 和 John Sweller。1998。《专业水平层级与教学设计》。《人因：人因与工效学学会杂志》40，1（1998年3月），1–17。doi:10.1518/001872098779480587

[24] Fatih Karataş、Faramarz Yaşar Abedi、Filiz Ozek Gunyel、Derya Karadeniz 和 Yasemin Kuzgun。2024。《将人工智能融入外语教育：ChatGPT 对外语学习者影响的研究》。《教育与信息技术》29，15（2024），19343–19366。doi:10.1007/s10639-024-12574-6 [25] Majeed Kazemitabaar、Runlong Ye、Xiaoning Wang、Austin Zachary Henley、Paul Denny、Michelle Craig 和 Tovi Grossman。2024。《CodeAid：评估一种兼顾学生与教育者需求的基于大语言模型的编程助手在课堂中的部署》。载于《2024年CHI计算机系统人为因素会议论文集》。计算机协会，纽约，纽约州，美国，文章650，20页。doi:10.1145/3613904.3642773[26] ChanMin Kim 和 Reinhard Pekrun。2014。《学习与表现中的情绪与动机》。载于《教育传播与技术研究手册》。施普林格，纽约，纽约州，美国，65–75。doi:10.1007/978-1-4614-3185-5_6[27] Kenneth R. Koedinger 和 Vincent Aleven。2007。《在认知导师系统实验中探索辅助困境》。《教育心理学评论》19（2007），239–264。doi:10.1007/s10648-007-9049-0[28] Shalece Kohnke 和 Tiffanie Zaugg。2025。《人工智能：促进STEM教育公平与机会获取的未开发机遇》。《教育科学》15，1（2025），68。doi:10.3390/educsci15010068[29] B. Kort、R. Reilly 和 R.W. Picard。2001。《情绪与学习相互作用的情感模型：重构教育教学法——构建学习伴侣》。载于《IEEE高级学习技术国际会议论文集》。IEEE计算机学会，洛斯阿拉莫斯，加利福尼亚州，美国，43–46。doi:10.1109/icalt.2001.943850[30] Bettina Laugwitz、Theo Held 和 Martin Schrepp。2008。《用户体验问卷的构建与评估》。载于《教育与工作中的人机交互与可用性》，Andreas Holzinger（编）。《计算机科学讲义》，第5298卷。施普林格，柏林、海德堡，63–76。doi:10.1007/978-3- 540-89350-9 6_[31] Mingyuan Li、Duan Wang、Erick Purwanto、Thomas Selig、Qing Zhang 和 Hai-Ning Liang。2025。《VisualCodeMOOC：一种整合对话代理的算法与数据结构课程平台，通过动态可视化增强学习》。《SoftwareX》30（2025），102072。doi:10.1016/j.softx.2025.102072[32] Irene Lopatovska 和 Ioannis Arapakis。2011。《图书馆与信息科学、信息检索和人机交互领域中情绪的理论、方法与当前研究》。《信息处理与管理》47，4（2011年7月），575–592。doi:10.1016/j.ipm.2010.09.001[33] Ewa Luger 和 Abigail Sellen。2016。《“ 就像有一个非常糟糕的私人助理 ”：用户对对话代理的期望与体验之间的鸿沟》。载于《 2016 年 CHI 计算机系统人为因素会议论文集》。计算机协会，纽约，纽约州，美国，5286–5297。doi:10.1145/2858036.2858288[34] Mehdi Malekzadeh、Mumtaz Begum Mustafa 和 Adel Lahsasna。2015。《智能辅导系统中的情绪调节综述》。《教育技术与社会》18，4（2015），435–445。jstor:jeductechsoci.18.4.435 https://www.jstor.org/stable/jeductechsoci.18.4.435[35] Richard E. Mayer。2002。《多媒体学习》。载于《学习与动机心理学》，第41卷。学术出版社，圣迭戈，加利福尼亚州，美国，85–139。doi:10.1016/s0079-7421(02)80005-6[36] Richard E. Mayer 和 Roxana Moreno。2003。《降低多媒体学习认知负荷的九种方法》。《教育心理学家》38，1（2003），43–52。doi:10.1207/S15326985EP3801 6_[37] Verónica-Alexandra Melo-López、Andrea Basantes-Andrade、Carla-Belén Gudiño-Mejía 和 Evelyn Hernández-Martínez。2025。《人工智能对融合教育的影响：系统性文献综述》。《教育科学》15，5（2025），539。doi:10.3390/educsci15050539[38] Roxana Moreno 和 Richard Mayer。2007。《交互式多模态学习环境》。《教育心理学评论》19，3（2007年6月），309–326。doi:10.1007/s10648-007-9047-2[39] Elham Mousavinasab、Nahid Zarifsanaiey、Sharareh R. Niakan Kalhori、Mahnaz Rakhshan、Leila Keikha 和 Marjan Ghazi Saeedi。2021。《智能辅导系统：特征、应用与评估方法的系统性综述》。《交互式学习环境》29，1（2021），142–163。doi:10.1080/10494820.2018.1558257[40] Thomas L. Naps、Guido Rößling、Vicki Almstrum、Wanda Dann、Rudolf Fleischer、Chris Hundhausen、Ari Korhonen、Lauri Malmi、Myles McNally、Susan Rodger 和 J. Ángel Velázquez-Iturbide。2003。《探索可视化与参与在计算机科学教育中的作用》。《ACM SIGCSE公报》35，2（2003年6月），131–152。doi:10.1145/782941.782998[41] Jum C. Nunnally 和 Ira H. Bernstein。1994。《心理测量学理论》（第3版）。麦格劳-希尔，纽约。[42] Heather L. O’Brien 和 Elaine G. Toms。2008。《什么是用户参与？定义用户与技术互动中用户参与的概念框架》。《美国信息科学与技术学会杂志》59，6（2008年2月），938–955。doi:10.1002/asi.20801[43] Wanda J. Orlikowski 和 Susan V. Scott。2008。《社会物质性：挑战技术、工作与组织之间的分离》。《管理学年鉴》2，1（2008），433–474。doi:10.5465/19416520802211644[44] Reinhard Pekrun。2006。《成就情绪控制—价值理论：假设、推论及其对教育研究与实践的启示》。《教育心理学评论》18（2006），315–341。doi:10.1007/s10648-006-9029-9[45] Reinhard Pekrun、Thomas Goetz、Wolfram Titz 和 Raymond P. Perry。2002。《学生自我调节学习与成就中的学业情绪：定性与定量研究计划》。《教育心理学家》37，2（2002年1月），91–105。doi:10.1207/s15326985ep3702_4[46] Sadhana Puntambekar 和 Roland Hubscher。2005。《在复杂学习环境中为学生提供支架式支持的工具：我们获得了什么，又错过了什么？ 》。《教育心理学家》40，1（2005年3月），1–12。doi:10.1207/s15326985ep4001_1[47] A mon Rapp、Lorenzo Curti 和 Arianna Boldi。2021。《人类—聊天机器人交互的人性面：关于基于文本的聊天机器人十年研究的系统性文献综述》。《国际人机交互研究杂志》151（2021年7月），102630。doi:10.1016/j.ijhcs.2021.102630

[48] Martina A. Rau。2017年。多种视觉表征提升STEM 学习有效性的条件。《教育心理学评论》29（2017），717–761。doi:10.1007/s10648-016-9365-3[49] Rolf Reber、Norbert Schwarz和Piotr Winkielman。2004年。加工流畅性与审美愉悦：美是否存在于感知者的加工体验之中？《人格与社会心理学评论》8，4（2004年11月），364–382。doi:10.1207/s15327957pspr0804_3[50] Roxana Rebolledo Font de la Vall和Fabián González Araya。2023年。探索人工智能语言学习工具的益处与挑战。《国际社会科学与人文发明期刊》10，1（2023），7569–7576。doi:10.18535/ijsshi/v10i01.02[51] Justin Reich和Mizuko Ito。2017年。从良好意图到真实成果：学习技术中的设计公平。技术报告。数字媒体与学习研究中心，尔湾，加利福尼亚州。https://tsl.mit.edu/research/from-good-intentions-to-real-outcomes-2/[52] Maryam Roshanaei、Hanna Olivares和Rafael Rangel Lopez。2023年。利用人工智能促进教育公平：机遇、挑战与新兴策略。《智能学习系统与应用期刊》15，4（2023），123–143。doi:10.4236/jilsa.2023.154009[53] Norsaremah Salleh、Emilia Mendes和John Grundy。2011年。高等教育计算机科学/软件工程教学中的结对编程实证研究：系统性文献综述。《IEEE 软件工程汇刊》37，4（2011年7月），509–525。doi:10.1109/tse.2010.59[54] David Scott和Eleanore Hargreaves（编）。2015年。《SAGE学习手册》。SAGE 出版社，伦敦。[55] Jeong-Bae Son、Natasha Kathleen Ružić和Andrew Philpott。2025年。用于语言学习与教学的人工智能技术及应用。《中国计算机辅助语言学习期刊》5，1（2025），94–112。doi:10.1515/jccall-2023-0015[56] Jan-Willem Strijbos和Maarten F. De Laat。2010年。构建计算机支持的协作学习中的角色概念：探索性综合。《人类行为中的计算机》26，4（2010），495–505。doi:10.1016/j.chb.2009.08.014[57] John Sweller。1994年。认知负荷理论、学习困难与教学设计。《学习与教学》4，4（1994年1月），295–312。doi:10.1016/0959-4752(94)90003-5[58] John Sweller、Jeroen J. G. van Merrienboer和Fred G. W. C. Paas。1998年。认知架构与教学设计。《教育心理学评论》10，3（1998年9月），251–296。doi:10.1023/a:1022193728205[59] N Tractinsky、A.S Katz和D Ikar。2000年。美即是可用。《与计算机交互》13，2（2000年12月），127–145。doi:10.1016/s0953- 5438(00)00031-x[60] Thomas Tullis和William Albert。2013年。测量用户体验：收集、分析和呈现可用性指标（第2版）。Morgan Kaufmann，沃尔瑟姆，马萨诸塞州，美国。doi:10.1016/C2011-0-00016-9[61] Viswanath Venkatesh和Fred D. Davis。2000年。技术接受模型的理论扩展：四项纵向实地研究。《管理科学》46，2（2000年2月），186–204。doi:10.1287/mnsc.46.2.186.11926[62] Lev S. Vygotsky。1978年。《思维与社会：高级心理过程的发展》。哈佛大学出版社，剑桥，马萨诸塞州。由Michael Cole、Vera John-Steiner、Sylvia Scribner和Ellen Souberman编。[63] Mark Warschauer和Tina Matuchniak。2010年。新技术与数字世界：分析获取、使用与成果公平的证据。《教育研究综述》34，1（2010年3月），179–225。doi:10.3102/0091732X09349791[64] David Wood、Jerome S. Bruner和Gail Ross。1976年。辅导在问题解决中的作用。《儿童心理学与精神病学杂志》17，2（1976），89–100。doi:10.1111/j.1469-7610.1976.tb00381.x[65] Franz Wortha、Roger Azevedo、Michelle Taub和Susanne Narciss。2019年。数字化学习环境中的多种负面情绪：来自两种方法的证据及其对学习的不利影响。《心理学前沿》10（2019），2678。doi:10.3389/fpsyg.2019.02678[66] Xi Yang、Marco Aurisicchio和Weston Baxter。2019年。理解与对话代理互动时的情感体验。载于《2019年CHI计算机系统人为因素会议论文集》。ACM，纽约，纽约州，美国，1–12。doi:10.1145/3290605.3300772[67] Ramazan Yilmaz和Fatma Gizem Karaoglan Yilmaz。2023年。使用基于生成式人工智能（AI）的工具对学生计算思维技能、编程自我效能和动机的影响。《计算机与教育：人工智能》4（2023），100147。doi:10.1016/j.caeai.2023.100147