3.1 引言

生成式人工智能系统，尤其是以 ChatGPT、Claude 与 Gemini 为代表的大型语言模型，正在深刻改变人们获取信息与开展知识性工作的方式。以 2022 年 11 月上线的 ChatGPT 为例，其在两个月内即获得一亿用户，成为历史上增长最快的消费级应用（Hu，2023）。到 2024 年，据报告 64% 的大学生与 58% 的知识工作者在学习、写作、编程与问题求解中经常使用生成式人工智能（Pew Research Center，2024；UNESCO，2023）。如此迅速的普及引发教育界、政策制定者与学术界的广泛讨论：生成式人工智能究竟是通过降低认知负荷与扩展专业知识可及性来促进学习，还是因诱发被动依赖而削弱学习并导致能力退化（Kasneci 等，2023；Mollick 与 Mollick，2023）。

现有关于人工智能教育应用的研究，主要关注任务层面的结果变量，包括准确性、效率与绩效指标。相关证据显示，人工智能辅助能够提升写作质量（Dang 等，2023）、加速编程任务（Kalliamvakou 等，2022），并促进创意构思（Gero 与 Chilton，2019）。然而，此类研究多衡量“借助人工智能产出了什么”，而较少揭示“学习过程中如何与人工智能进行认知层面的互动”。这种以产出为中心的取向掩盖了关于学习过程的关键问题：学习者是在主动建构知识，还是在被动接受生成内容；他们是否监控自身理解、评估人工智能的可靠性并调节依赖程度；哪些元认知策略区分了有效学习者与因过度依赖而损失能力的学习者。

与之并行的人机交互研究强调信任校准与恰当依赖，指出用户既可能出现过度信任（接受错误输出），也可能表现为不足信任（拒绝正确输出）（Glikson 与 Woolley，2020；Lee 与 See，2004；Bansal 等，2021；Zhang、Liao 与 Bellamy，2020）。这些研究通过揭示自动化偏差（Goddard、Roudsari 与 Wyatt，2012）与算法规避（Dietvorst、Simmons 与 Massey，2015）等认知偏差做出了重要贡献。但相关文献往往将用户视为在专业水平类别内较为同质的群体，假定新手普遍更易过度信任，而专家更能保持适度怀疑。此一理论图景常以“专家—新手”的二分范式呈现，并据此提出设计建议：为初学者提供更多支架，为专家提供更多自主性（Amershi 等，2019；Kocielnik、Amershi 与 Bennett，2019）。

值得注意的是，近期的实证观察对这一专家—新手范式提出挑战。教育者发现，即便在同一水平层次内，学生的人工智能使用质量差异显著：有的博士生提交“带有 ChatGPT 风格”的论文，语言流畅但逻辑松散；而部分本科生却展现出对人工智能输出的较高层次批判性参与（Rudolph、Tan 与 Tan，2023；Sullivan、Kelly 与 McLaughlan，

2023）。产业一线亦报告，具有相近从业年限的同事在人机协作策略上差异明显，有人严谨核验每一项论断，也有人近乎无条件信任输出（Prunkl 等，2021）。这些观察表明，仅以专业水平并不足以预测人工智能使用的有效性。若传统的用户特征（受教育程度、领域知识与工作经验）不足以解释效果差异，那么关键的解释变量何在。

本研究提出，遗漏的关键因素在于元认知，即个体在与人工智能协作时对自身认知过程的意识与调节（Flavell，1979；Schraw 与 Dennison，1994）。元认知涵盖规划（任务分解、目标设定、策略选择）、监控（理解追踪、质量评估）、评价（能力判断、风险识别）与调节（方法调整、工具切换、依赖控制）。在传统学习情境中，元认知能力对学业成就具有独立于智力与先验知识的强预测力（Veenman、Van Hout-Wolters 与 Afflerbach，2006；Zimmerman，2002）。能够主动监控理解、评估信息源可靠性并调节学习策略的学生，往往优于在领域知识上相当但元认知能力较弱的同伴（Winne 与 Hadwin，1998）。

尽管元认知在学习科学中居于核心地位，但在人工智能教育应用研究中仍相对缺席。本文对 2020—2024 年间 127 篇教育类人工智能研究的系统性梳理显示，仅有约 7% 的研究明确考察用户的元认知过程，且尚无研究探讨元认知策略如何塑造人机协作的有效性（作者自述之综述）。这一缺口尤为重要，因为生成式人工智能改变了认知工作的重心。相较于信息检索类工具，生成式人工智能能够直接执行分析、综合、推理与创作等“认知劳动”。因此，用户必须在互动过程中持续作出元认知判断，包括何时将认知任务委派给人工智能，何时保持人类主导，如何实施核验，以及如何防止能力退化。当前缺乏来自真实情境的经验证据来揭示这些元认知实践的具体形态、跨个体差异以及系统设计如何有效支撑。

上述缺口对理论与实践均具有重要影响。在理论层面，缺乏元认知视角限制了我们解释“为何背景相似的用户会做出截然不同的人工智能使用决策”。既有强调信任（Lee 与 See，2004）、技术接受（Venkatesh 等，2003）或专业水平阶段（Dreyfus 与 Dreyfus，1986）的框架，均难以充分解释群体内部的显著差异。在实践层面，误将使用成效归因于“专业水平”，而非“元认知策略”，可能导致错误的适应式系统设计，将支架提供给不恰当的用户群体，甚至强化而非纠正问题性使用模式。本研究采用设计科学研究（Design Science Research, DSR）范式,该范式通过关注创造和评估能够扩展人类和组织能力的创新性人工制品，从而区别于传统的行为科学研究（Hevner et al., 2004）。行为科学寻求发展和验证解释或预测现象的理论，而设计科学则通过系统化

地构建和评估有目的的人工制品来解决实用性问题。在我们的研究背景下，这意味着我们不仅仅寻求理解用户当前如何与 AI 系统交互，而是要生成关于 AI 系统应该如何设计以支持元认知参与的规范性知识。

遵循 Gregor 和 Hevner（2013）提出的 DSR 知识贡献框架，我们将本研究定位于改进（Improvement）和移用（Exaptation）象限的交叉点。这一定位反映了我们贡献的双重性质。一方面,我们基于成熟的 AI 教育系统领域，通过纳入元认知支持机制来寻求改进其设计。这代表了改进维度，其中问题空间（教育中的 AI）和解决方案空间（自适应学习系统）都已达到相当成熟度。另一方面，我们应对支持用户在 AI 中介学习情境中的元认知这一新兴挑战，这代表了一个具有有限现有解决方案的新应用情境。这一移用维度承认，虽然 AI 教育工具已经很成熟，但将它们适配以专门促进元认知发展而不仅仅是传递内容，代表了一种新的问题表述。我们的研究在多个抽象层次上贡献知识，遵循 Gregor 和 Hevner（2013）的人工制品分类框架。在构念（construct）层面，我们发展了新的概念词汇，包括元认知复杂度评分系统和六类用户模式类型学。在模型（model）层面，我们提出了一个整合框架，将这六种元认知使用模式（标记为 A 至 F）映射到特定的设计需求。在方法（methods）层面，我们贡献了一种通过交互分析识别和分类用户元认知模式的系统方法。最后，在理论（theory）层面，我们推进了一种新生设计理论，提出 AI 系统应该基于检测到的元认知模式而非静态用户特征（如专业水平或人口统计属性）来调整其支持策略。这种多层次的贡献结构既能在系统设计中立即实际应用，又能在人机协作研究中实现长期理论进步。

基于上述理论与实践问题，本研究聚焦不同用户在学习与知识性工作中与生成式人工智能协作时呈现的元认知使用模式。我们并不预设“专家—新手”“理工—人文”等先验分类，而是采用扎根理论方法，从经验材料中归纳模式（Charmaz，2014；Glaser 与 Strauss，1967）。研究共开展 49 场半结构式深度访谈（每场 45—93 分钟），采用最大差异取样策略，覆盖本科生至资深从业者，涵盖计算机科学至人文学科等多种背景，并兼顾从日常高频使用者到偶尔使用者的频率差异。借助回顾式口语报告分析（Ericsson 与 Simon，1993）与关键事件技术（Flanagan，1954），引导受访者在细粒度层面重构近期的人机互动，并重点追问其规划、监控、评价与调节过程。

本章节围绕以下三个问题展开。研究问题 1：在学习与知识性工作情境中，不同用户与生成式人工智能协作时呈现出哪些元认知使用模式。该问题旨在识别用户所采用的相对稳定的元认知策略组合，模式并非预设与人口统计属性对齐，而是从行为

与认知数据中归纳而出。研究问题 2：这些使用模式与用户特征（专业水平、学科背景、经验）及情境因素（任务重要性、领域熟悉度、时间压力）之间呈现何种关系。该问题考察传统用户分类能否预测元认知模式，抑或模式跨越人口学边界，同时关注同一用户在不同情境下的策略变动。研究问题 3：由有效与无效的元认知模式可推导出何种系统设计需求。该问题旨在将实证发现转化为可操作的设计原则，以促进促进学习的协作而非削弱学习的协作。

研究结果显示，存在六类具有代表性的元认知使用模式，对人工智能采纳的既有认知提出根本性挑战。在直接观察样本（N=49）中，五种主要模式呈现显著差异化特征。模式 A（战略性分解与控制，37%）通过主动的任务分解维持严格的人类主导。模式 C（情境敏感的适配，33%）依据任务属性动态校准信任与介入深度。模式 E（教学化反思与自我监控，14%）将人工智能作为促进元认知觉察的反思工具。模式 B（迭代优化与校准，8%）与模式 D（深度核验与批判性介入，8%）分别通过容错迭代和系统化验证实现有效协作。此外，模式 F（无效与被动使用）虽在质性分析和教师访谈中被识别为第六种模式，但在本研究样本中未有受访者将其作为主要策略。基于教师课堂观察的非正式估计提示，约 25-40% 的学生群体可能呈现该模式特征，表现为无批判地接受输出、缺乏过程监控以及依赖程度的无意识加深。

本研究的贡献主要体现在三个方面。第一，在理论上，提出一个基于模式的人机协作解释框架，超越传统的专家—新手二分。通过识别六种以元认知为基础的使用模式，并显示其跨越人口学边界，该框架为“背景相似但结果迥异”的现象提供了更为细致且经验证据支持的解释。同时，本文将交互记忆系统理论拓展至人—人工智能二元协作，指出有效协作并非机会主义的认知外包，而是即便牺牲效率也要坚持的边界维护原则，此一发现对组织协调与知识管理研究亦具启发意义（Wegner，1987）。第二，在方法上，以元认知过程为分析单元提出研究路径创新。相较于既往以外显行为（点击、用时）或事后结果（准确率、满意度）为主的研究，本文采用口语报告与关键事件技术捕捉通常不可见的认知过程，构建并在 49 名受访者中验证了包含 12 个子过程的元认知分类体系，为后续研究提供可复用的测量与编码框架。第三，在实践与设计上，提出面向“元认知协作代理”（Metacognitive Collaborative Agents, MCA）的 19 项基于证据的设计需求。这些需求对强调无缝自动化与顺滑对话的既有设计理念提出修正。例如，需求 MR13（不确定性透明呈现）要求系统显式传达置信程度并标注知识边界，从而纠正当前系统语言流畅但认知过度自信的倾向；需求 MR16（防止技能

退化）要求监测用户的独立性比率，并在过度依赖威胁能力保持时实施干预。上述需求为下一代教育类人工智能系统的研发与学习关键场景的部署提供了可操作的规范。

3.2 理论基础

生成式人工智能工具在教育领域的迅速普及，引发了关于其有效性的广泛研究。然而，一个关键维度仍然相对不足，即学习者在学习过程中的认知参与方式。既有研究多集中于结果层面，如准确性提升、效率改善与任务完成率，但较少关注中介这些结果的元认知过程。本文献综述旨在为考察人机协作中的元认知使用模式奠定理论基础，认为以专业水平划分与静态信任模型为根基的传统方法，难以充分解释人工智能有效性的大幅异质性。通过梳理元认知理论、信任框架、人工智能采用模式及其理论缺口，本文主张在学习情境下区分有效与无效的人机协作，应以元认知视角为核心。

3.2.1 元认知理论与自我调节学习

理解学习者如何思考其自身的思维活动，即元认知，为有效学习策略提供基础性洞见。相关研究历经半个世纪，已表明元认知能够独立于智力或先验知识而显著预测学习成效。其在人与人工智能协作中的应用仍处于起步阶段。Flavell（1979）系统开创了元认知研究，将其界定为“关于认知现象的知识与认知”（第 906 页）。其框架区分并强调四个相互作用的组成：元认知知识（学习者对认知的认知）、元认知体验（任务过程中的有意识认知或情感反应）、元认知目标（激活元认知活动的目标）与元认知行动（为实现目标而采取的策略）。该架构将元认知置于监控与控制低层次认知过程的高阶系统位置。进一步地，Flavell（1985）将元认知凝练为深刻影响后续研究的两个要素：认知的知识（元认知意识）与认知的调节（元认知控制）。Wang 等（1990）的综述得出结论，元认知是普遍意义上学习绩效的最重要预测因子，确立了其在教育心理学中的核心地位。这种“知”与“行”的区分，即理解自身认知过程与主动调控它们之间的差异，对于解释人机协作尤为关键。学习者可能知道人工智能输出需要核验（元认知知识），但在实践中并未落实相应的核验策略（元认知调节）。在以人工智能介入的学习环境中，由于生成内容的便捷性，调节过程更易被绕过，知识与行动的鸿沟尤为突出。

对于元认知过程的测量与操作化，Schraw 与 Dennison（1994）以“元认知意识量表”（MAI，52 题）将 Flavell 的框架操作化，测量成人元认知意识的两个主要因子。在对 307 名本科生的实证中，两个因子的内部一致性均较高（$\alpha=0.90$），并与学习表

现呈显著相关。研究进一步明确了四类关键的元认知调节过程：规划（任务前策略选择与资源分配）、监控（任务中对理解与策略效果的持续评估）、评价（任务后回顾性评估）与调节（为促进学习而作出的调整）。Young 与 Fry（2008）的研究显示，MAI 得分与累计绩点显著相关（r＝0.23，p＜0.01），且研究生在“认知调节”上的得分显著高于本科生（M＝142.04 vs. 136.85，F(1,177)＝4.13，p＜0.05）。调节因子对高级学习者的重要性提示：关于认知的知识与对认知的控制存在相对独立的发展路径。在人机协作中，这一发现具有重要启示意义，学习者必须主动调节对外部工具的依赖。MAI 的分维有助于厘清与人机协作密切相关的具体能力。陈述性知识（“我了解自己的智力优势与不足”）影响何时适宜求助于人工智能。程序性知识（“我知道如何有效使用策略”）支撑高水平的提示工程与核验策略。条件性知识（“我知道在何时何因使用何种策略”）有助于依赖的恰当校准，即辨识哪些任务适合借助人工智能，哪些应自主完成。然而，具备这些知识与在实践中真实运用之间常存在落差，单纯的量表难以充分捕捉。

而 Winne 与 Hadwin（1998）提出四阶段自我调节学习模型，强调学习的递归与动态过程。学习者经历任务界定（解读任务与认知条件）、目标与规划（设定目标与策略路径）、学习策略实施（策略执行）、元认知适应（对未来表现的较持久认知条件改变）。各阶段依托 COPES 认知架构：条件（资源与约束）、操作（检索、监控、整合、复述、转化）、产出（结果）、评估（执行中的判断）、标准（成功的准则）。监控与控制是各阶段的枢纽，支持学习者在将产出与标准比较的过程中持续反馈与更新。Greene 与 Azevedo（2007）基于 113 篇文献的评述指出，该模型通过承认学习的连续调整而非线性推进，能够解释以往研究中的矛盾发现。在以人工智能介入的学习中，这种递归结构尤为贴切。学习者可能先将大量工作委托给人工智能（任务界定），在监控中发现不利于深度理解（评估），因而转而将人工智能作为核验而非生成的工具（调节），并将这一改变内化用于后续任务（元认知适应）。然而，现有研究鲜少捕捉这些动态调整，更多关注稳定的使用模式。

伴随着技术介入的元认知支持，Azevedo 与 Hadwin（2005）将自我调节学习理论延展至技术增强的学习环境，展示了计算机系统在支撑元认知过程上的有效性。其综述考察了书面提示、静态支架与自适应支持等多种方式对陈述性、程序性、概念性与元认知知识的影响。训练研究表明，接受 SRL 策略教学的学生在超媒体学习中表现更佳（Azevedo & Cromley，2004），对技术整合的 SRL 策略持积极态度者从事自我反

思与适应行为的可能性提升 45%。MetaTutor 系列研究显示，充当外部调节者的教学代理能够有效支架认知与元认知策略，且通过多模态数据（日志、眼动、表情）可实时捕捉认知、情感、元认知与动机的展开（Azevedo 等，2013）。上述证据表明，合理设计的技术并不必然削弱元认知，反而可以支撑之。然而，生成式人工智能与超媒体支架环境存在本质差异。前者直接产出成品，可能缩短促成学习的“必要磨砺”。学习者在生成式人工智能情境中是否沿用相同的元认知调节策略，仍有待经验检验。

目前，已有相关研究将元认知能力作为学习结果的独特预测因子。Veenman 等（2006）的综合研究表明，元认知技巧在控制智力之后仍能显著预测学习表现，从而回应了“元认知是否仅是一般智力的另一面向”的理论疑问。其整合结果显示，智力对学习绩效的独特解释约为 10%，元认知技巧的独特解释为 17%至 18%，两者共有方差为 20%至 22%。智力与元认知的相关平均在 0.40 至 0.45 之间，说明二者相关而相异。重要的是，元认知在智力之外仍能预测学习成效，这意味着足够水平的元认知可在一定程度上弥补较低智力所带来的不利。这一结论对人工智能介入的学习具有深远意义。如果元认知能力与乃至超过领域专长同等关键，那么简单的专家与新手划分难以预测有效的人机协作。Van der Stel 与 Veenman（2008，2010）从发展角度延伸了上述结论。在控制智力后，13 至 15 岁样本中元认知技巧质量均能显著预测数学成绩，且预测力随年龄提升。Zimmerman（2002）提出循环三阶段模型，将元认知与动机过程整合为事前（任务分析与自我效能）、执行（自我控制与自我观察）与自我反思（自我评判与自我反应）。Kitsantas 与 Zimmerman（2002）的微观分析测量显示，自我调节指标可解释超过 90%的技能绩效方差，佐证元认知过程的强预测效力。总体而言，能够有效监控认知、评估理解并调节策略的学习者，较高智力但低元认知者更具优势。

也有相关研究将元认知划分为知识与调节，这不仅是分类学意义的问题，更是刻画了学习中的基本张力。Schraw 与 Moshman（1995）将元认知知识细分为陈述性（例如“我难以记住日期”）、程序性（例如“我会使用概念图”）与条件性（例如“历史适合摘要，数学问题则未必”）。相应地，元认知调节涵盖学习中的规划、监控与评价等动态过程。尽管二者在实证上相关（多项研究相关在 0.54 至 0.73 之间），但仍可区分。研究生在“调节”上的优势并不必然伴随更高的“知识”（Young & Fry，2008），提示调节更依赖有意识练习，而非被动积累。这一点与 Winne（1996）和 Veenman 等（2006）的强调一致：具备元认知知识并不保证其应用。学习者可能明知人工智能输出需核验，却在执行时未予落实。该“知—行”缺口很可能随个体与情境而变化，传统

的专业程度划分难以捕捉。因而，有必要考察学习者在与人工智能协作时如何实际开展元认知调节，这正是当前研究所匮乏的过程视角。然而，现有元认知理论主要提供了描述性和解释性知识，关于如何将这些理论转化为 AI 系统的具体设计原则与支持机制，仍缺乏系统性的规范性研究。

3.2.2 人机协作中的信任与依赖

元认知理论阐明学习者如何调节自身的认知，信任框架则关注学习者在何时依赖外部代理（包括人或自动化系统）。尽管现有信任研究在理论上已相当丰富，但仍存在明显局限：往往将用户视为在专业类别内近乎同质，并将信任概念化为相对稳定的个体层面特质，而非动态、情境依赖的过程。Lee 与 See（2004）提出了自动化信任的基础框架，将信任定义为“在不确定与脆弱情境下，相信某代理将有助于实现个体目标的态度”（第 54 页）。其核心见解在于，信任的关键是校准问题。过度信任导致误用（缺乏监控的依赖），而不足信任导致弃用（拒绝有用的自动化）。恰当依赖来自与系统真实能力相匹配的信任校准。作者据此提出信任恰当性的三个维度：分辨率（信任对能力差异的辨识精度）、特异性（信任能否针对特定部件与时间变化）与校准本身（信任与能力的最终对应）。信任形成源于分析性、类比性与情感性三类过程。该框架强调自动化依赖是自我信心与对系统信任之间的权衡。当信任超过自我信心时，依赖增加；当自我信心占优时，倾向独立完成任务。这一点对于学习情境尤为重要，因为自我信心随任务难度与领域熟悉度而大幅波动。困境中的学生可能适度求助于人工智能，而过度自信者可能在具备独立完成能力时仍不当依赖。需要指出的是，该框架主要针对能力相对稳定的自动化系统，对于性能可变、可学习的自适应人工智能系统下的动态信任调整关注不足。

而对于算法依赖的系统性偏差方面，两类互补研究揭示了对恰当依赖的系统性偏离：自动化偏差（过度依赖）与算法厌恶（不足依赖）。Goddard 等（2012）对医学决策支持中的自动化偏差进行系统综述，认为用户将计算机输出视为“对警觉性信息搜寻的启发式替代”（第 121 页）。其元分析显示，错误建议使错误决策风险上升 26%（风险比＝1.26，95%CI：1.11—1.44），6%至 11%的案例出现“负面咨询”，即原本正确的判断因建议而转为错误。然而，由于“正面咨询”（12%至21%）多于“负面咨询”，净效应往往仍为正（总体改善 6%至 8%）。经验的作用较为复杂：任务经验不足会增加自动化偏差，而对特定系统的熟悉有时反因过度熟悉而滋生自满。负荷、时间压力与任务复杂性均会通过挤压认知资源而提升依赖。相反，Dietvorst 等（2015）揭示了

算法厌恶，即在观察到算法错误后，个体倾向于不合理地回避算法，即便算法明显优于人类预测者。在五项实验中，相比于人类犯同样的错，参与者对算法的信心下降速度快 2 至 3 倍，随后更倾向选择较差的人类预测者。其机制在于，人们对算法抱有近乎完美主义的期待，却能容忍人类的易错性。后续研究发现，允许用户对算法输出进行微小调整（±10 分）即可显著提升算法采用，说明“感知到的可控性”比“客观优越性”更能维持恰当依赖（Dietvorst 等，2016，2018）。述相反的模式表明，信任与依赖高度依赖任务性质、性能透明度与先验期待。多数研究却将这些变量视为稳定倾向，而非对具体情境的动态反应。在教育场景中，同一名学生可能在被视为客观的技术型任务上出现自动化偏差，在主观写作或创造性任务上表现出算法厌恶，且两者可在同一学习时段内交替出现。

Glikson 与 Woolley（2020）对人机信任文献进行综合评述，从人工智能的呈现形式（嵌入式、虚拟体、机器人）与机器智能水平两维展开。其识别了广泛适用于各类人工智能的认知性信任因素（可感知性、透明性、可靠性、即时行为）与情感性信任因素（拟人化与对类人特征的社会性反应）。评述指出，人际信任通常随互动时间增加而上升，而技术信任可能因遇到边界案例与失败而随时间下降。人工智能具备学习、演化与主动行为的能力，这些行为可能难以解释，区别于传统自动化，从而带来新的信任挑战。评述同时强调，任务类型、组织情境、感知风险与不确定性等因素均会调节显著的信任路径。然而，重要局限在于：研究往往将专业类别内的用户视为同质，假定专家能恰当校准信任，而新手的过度或不足依赖主要源于领域知识不足。最新证据挑战这一假定（Schemmer 等，2024），显示心理特质如信任倾向、需要认知与自我效能会在专业类别之内产生显著差异。个体的认知风格、人格特质（尤其是开放性、尽责性与神经质）与任务熟悉度（不同于领域专长）均可在专业程度之外预测信任与依赖。相关研究表明，这些心理因素可在信任之外解释高达 24%的依赖行为方差，提示以专家或新手为单位的同质性假设掩盖了关键的个体差异。

可解释人工智能（XAI）的兴起基于一个假设：透明性能够促进恰当依赖。然实证证据显示其与信任之间关系更为复杂，甚至悖反。Bansal 等（2021）总结了“令人困惑的研究图景”，指出解释对信任与团队绩效的影响并不一致。最准确的人工智能并不必然带来最佳的人机团队表现；解释有时提升信任，却也会增加“盲目信任”，即对正确与错误建议的同时过度依赖。研究表明，仅提供置信度信息效果有限，“数据可得性”式解释会在用户自我能力较低时诱发过度信任，“特征重要性”解释常显得有说服力，

但并不能真正支持核验。来自微软与 HCI 顶会的后续研究同样发现，详尽解释常常增加对错误建议的过度依赖，高保真解释使用户误以为人工智能的推理近似人类，从而触发拟人化，削弱批判性评估。即便是信息量不高的解释（如 50% 的准确率提示）也能提升信任，说明解释更像“说服”而非“核验”工具。复杂度亦关键：只有在需要大量努力的任务中，解释才更可能降低过度依赖，因为用户具备批判评估解释的动机与能力。将解释与“认知强迫”类干预相结合，即要求用户进行刻意评估的机制，较单纯解释显示出更大潜力（Vasconcelos 等，2023）。该悖论提示，仅披露人工智能的推理过程并不足以确保恰当依赖。用户不仅需要解释，更需要元认知意识以进行批判评估，识别自身知识边界，并监控人工智能输出是否契合其学习目标。XAI 文献由此侧面揭示了缺失，即关于用户在人工智能辅助决策中元认知过程的关注。

动态信任与个体层模型也存在一定的局限，既有信任框架多将信任概念化为较稳定的特质，无论是个体层面的信任倾向，还是系统层面的“可置信度”。最新证据显示，信任高度动态且情境依赖。其发展轨迹倾向不对称：相较于建立，信任的破坏速度快 2 至 3 倍；早期经历会形成持久锚定；一旦发生信任违背，即便后续表现良好，信任也难以完全恢复（Siau & Wang，2020）。任务属性（客观与主观、熟悉与新奇、高风险与低风险）会动态改变恰当的信任阈值。环境因素如负荷、时间压力与组织规范也会重置信任基线。人工智能性能的变异同样重要，而用户在遇到分布外数据时，往往维持对人工智能的高信任，同时降低对自我的信任，这恰是与恰当校准相反的方向（Zhang 等，2024）。新近研究承认“信任具有情境依赖性”，并指出“没有一种信任模型可适用于所有情境”（Lockey 等，2024），但将这一认识落地于操作层面的研究仍然稀缺。现有框架对实例层面的波动关注不足，如信任如何随具体推荐语境、近期交互历史、可见置信信号以及用户当下的疲劳与认知负荷而瞬时变化。上述动态很可能与元认知监控与调节相互作用：元认知较强的学习者更可能做出恰当的信任调整，元认知较弱者则可能在不同情境下持续保持不当的稳定信任水平。因而，有必要将信任视作任务执行中的动态过程，而非静态的个体特质或系统属性。值得注意的是，现有信任研究主要聚焦于描述和解释信任校准问题，但关于如何设计 AI 系统以主动支持用户进行恰当信任校准与元认知监控的设计知识仍然匮乏。

3.2.3 学习情境下的人工智能使用模式与采用

生成式人工智能在教育中的快速扩散催生了关于采用模式、生产率效应与风险关注的实证研究。然相关证据呈现一个悖论：即便在专业程度相近的人群中，使用效

果的差异仍然显著，提示传统类别划分难以解释成功的人机协作。高等教育对生成式人工智能的采用显著加速。学生使用比例从 2023 年春季的 27%升至当年秋季的 49%，至 2025 年达到 92%，其中 88%用于评估活动，而前一年仅 53%（Giray，2023；UNESCO，2023）。教师采用相对滞后，约 25%报告经常使用，对比学生的 49%；约 40%的教师自评仍处在“入门阶段”，仅 17%达到较高水平。Kasneci 等（2023）综合分析指出，人工智能在个性化学习、内容创作、自动化反馈与语言支持等方面具备机会，同时伴随偏见强化、幻觉、知识时效性与抄袭等挑战。其核心洞见是，学生需要扎实的背景知识以评估人工智能输出，这实质上是元认知能力的体现，而该能力在同一专业组内差异显著。Mollick 与 Mollick（2023）提出人工智能在教学中的七种角色：导师、教练、导师型支持、队友、工具、模拟器与学生。不同角色的优势与风险不同，且均要求学生保持“人在回路”地位。将人工智能作为导师时，需监控解释是否真正提升理解；作为队友时，需批判性评估并整合其贡献；作为工具时，需判断何种任务可适度委托。实践中，学生在这些元认知辨别上的表现差异巨大，不少学生即便在示范后仍“在提示中提供极少信息”，未能有效发挥人工智能能力。

相关研究也提供了生产率与学习结果的实证证据，首先是关于 GitHub Copilot 的随机对照实验提供了相对严格的生产率证据。Peng 等（2023）在 95 名专业程序员中进行试验（以 JavaScript 实现 HTTP 服务器），处理组完成任务时间较对照组缩短 55.8%（71.17 分钟对 160.89 分钟，95%CI：21%—89%，p＝0.0017）。多数参与者报告更易进入“心流”（73%）、更有满足感（60%—75%）、在重复性任务中保存心智资源（87%）。值得注意的是，自评生产率提升（35%）显著低估了实际提升，提示用户对人工智能效应的元认知准确度不足。更重要的异质性效应挑战了“专业放大”假设：经验较少者受益更多（系数 8.23，p＝0.0629），高强度编码者受益多于低强度者（系数 −11.70，p＝0.0168），25—44 岁组的提升多于更年轻者（系数 −74.55，p＝0.0303）。这表明人工智能具有“技能均衡器”的功能，而非仅仅放大已有专长。Kalliamvakou 等（2022）对 2047 名开发者的调查亦发现心理收益随生产率提升而显著，且“接受度”驱动感知生产率。其次，学习结果研究总体呈正效应。多项元分析报告学业成绩的积极效应值，部分研究显示考试成绩可提升约 10%（Baidoo-Anu & Owusu Ansah，2023；Chen 等，2025）。但平均效应掩盖了重要的异质性：采用“掌握取向”的学生（以构建与扩展知识为目标使用人工智能）学习提升显著高于“绩效取向”的学生（为快速完成任务而使用人工智能）。训练至关重要：接受人工智能训练的工程学

生在测评中得分 6.60，高于无训练但使用人工智能者的 4.94 与对照组的 4.28，并在布鲁姆分类的多个层次上改善（Kim 等，2025）。这些训练很可能提升了提示优化、输出质量监控与与学习目标对齐等元认知策略，尽管多数研究并未直接测量这些过程。

目前，对于专业类别内部差异的证据，至少五条证据线索挑战“专业水平预测人工智能使用效能”的假设。第一，Copilot 的异质性效应直接反驳“放大假设”，新手受益更多，提示人工智能更像“均衡器”。第二，人工智能素养研究显示，同一群体内部差异显著，技术背景与人工智能经验较学术年级与领域专长更能预测结果（Ng 等，2021；Strzelecki，2023）。第三，师生采用差距与预期相反，学生（领域新手）在速度与能力上均超过教师（领域专家），提示领域专长难以直接迁移为人机协作能力。第四，即便在同一课程内，提示工程的有效性差异巨大，基础模式对各层级有效，而高级模式对技术熟练者亦颇具挑战。第五，首代大学生在恰当使用情境上的自信不足，即便其学术能力相当，提示社会经济因素独立影响元认知与使用策略。Amershi 等（2019）基于 20 个微软产品、49 名实践者提出的人机交互 18 项设计指南，亦侧面支持这一批评。指南承认有效性更多取决于情境与任务属性，而非用户专业分类。尤其是第 11 条（澄清系统为何如此行为）在广泛 XAI 研究的投入下仍最易被违背，说明“解释”在没有元认知评估能力的前提下难以发挥核验功能。该指南组隐含假设用户群体相对同质且错误模式一致，而现实显示，元认知监控、认知风格与信任倾向等个体差异带来异质的使用模式，标准化指南难以充分覆盖。

生成式 AI 的出现，也带来了认知卸载、技能退化与过度依赖问题，Risko 与 Gilbert（2016）关于认知卸载的综述为“人工智能依赖”忧虑提供理论基础。认知卸载被界定为“通过身体行动改变任务的信息处理需求，以降低认知负荷”（第 676 页）。卸载短期有益，能够降低负荷并提升复杂任务表现，但长期可能因练习减少而导致技能退化。关键在于，卸载倾向依赖于元认知评估，而该评估“可能是错误的，从而导致次优卸载行为”（第 682 页）。学习者可能基于对任务难度、自身能力或努力加工学习价值的误判，将认知工作不当地转移给人工智能。新近研究提供了经验证据。Macnamara 等（2024）指出，较之传统自动化，人工智能辅助更可能加速技能退化，其过程大致经历探索、整合、依赖与依存四阶段。更令人担忧的是，这一退化常在“使用者不自觉”的情况下发生，反映为元认知失灵，即用户自认为保持了技能，但事实上已出现萎缩。针对 285 名学生的调查发现，68.9%因过度依赖而表现出“学业惰性”的提升（Rahiman & Kodikal，2023）。系统性综述亦指出，过度依赖削弱批判性思维、决

策与分析推理。“能力错觉”的概念揭示，人工智能生成的高质量外显产出可能误导个体对自身掌握程度的判断，从而掩盖深层认知缺失（Okonkwo & Ade-Ibijola，2023）。上述现象可被概括为“能动性衰减”，即独立思考与自主行动能力的逐步流失。对于具备较强元认知能力的学习者而言，这种风险或可减轻，因为其会主动监控理解、评估在无辅助条件下的解题能力，并策略性地调配人工智能的使用以维持技能发展。关键变量并非使用频率本身，而是学习者是否将其置于学习目标导向的元认知调节之下，而非仅仅服务于绩效目标。当前研究聚焦总体使用与结果，鲜少检视上述调节过程。

综合前述证据，基于专业水平的分类框架存在五方面不足。第一，Copilot 的数据反驳了放大假设，新手受益更多，人工智能具有“均衡器”而非“放大器”特征。第二，Amershi 的指南显示，有效性主要由任务情境决定，二元专业标签信息不足。第三，在人工智能素养、提示工程与使用成效中，组内差异常大于组间差异，技术背景与人工智能经验的重要性超过领域专长。第四，学生在采用与使用上超过教师，提示领域专长难以迁移为人工智能能力。第五，能力呈多维结构，如技术理解、实践应用、批判评估、元认知监控、提示工程与伦理推理等，个体可能在某些维度是“专家”，在另一些维度仍是“新手”，单维度分类因而失效。替代框架正在涌现。人工智能素养模型区分“知—理解”“用—应用”“评—创”“伦理—规范”四个相对独立的维度（Ng 等，2021）。交互模式框架关注接受率、核验行为、整合策略与元认知监控，而非静态的专业水平。能力画像将能力重构为“技术技能×领域知识×提示工程×批判评估”的矩阵，而非单一连续体。这些框架共同承认，基于行为的模式优于人口统计或专业类别，能更好预测有效性。然而，它们多停留在理论层面，缺少对用户在与人工智能协作中如何具体部署元认知策略的经验证据。更为关键的是，现有研究主要聚焦于描述和解释使用模式的异质性，但缺乏将这些实证发现转化为系统设计需求与元认知支持机制的规范性研究。如何构建能够检测、适应并促进用户元认知发展的 AI 教育系统，仍缺乏明确的设计原则与可操作的实施框架。

3.2.4 研究缺口与设计科学正当性

文献综述揭示了三个关键缺口，它们共同证明了采用设计科学研究方法进行本调查的合理性。首先，虽然 AI 过度依赖现象在近期研究中获得越来越多的认可（Cai et al., 2019; Goddard, Roudsari, & Wyatt, 2012; Parasuraman & Manzey, 2010），现有工作主要集中在记录问题而非开发系统化解决方案。我们拥有描述何时以及为何发生过度依赖的大量 Ω 知识，但缺乏规定如何设计通过元认知支持来预防或缓解这种过度依赖

的 AI 系统的 Λ 知识。当前方法倾向于关注改善 AI 可解释性或使用户能够提出异议，这些都是有价值的策略，但未能直接解决用户与 AI 系统的元认知参与。通过元认知支持能够培养更主动、批判性和反思性 AI 使用的具体机制在文献中仍然规定不足。

其次,现有 AI 教育系统在元认知支持功能方面表现出低解决方案成熟度。虽然自适应学习系统在基于学习者表现调整内容难度、排序和呈现方面已达到相当的复杂程度(Woolf, 2010; Vanlehn, 2011),但它们不会适应用户的元认知策略或主动支持元认知发展。当前的个性化方法通常依赖于基于专业水平、学习风格偏好或表现历史的静态用户模型。没有现有系统能够实时动态检测和响应用户的元认知参与模式。这代表了解决方案成熟度的根本缺口。我们有成熟的内容适应解决方案,但元认知适应解决方案不成熟。该领域缺乏展示如何在实践中操作化模式响应式元认知支持的已实施系统。第三,也是从设计科学角度最关键的,该领域缺乏关于如何设计支持而非取代元认知的 AI 系统的规范性 Λ 知识。虽然我们有关于脚手架、可解释性和可辩驳性的设计原则,但我们缺乏专门面向在 AI 交互情境中培养元认知参与的整合设计框架。现有设计知识解决了拼图的片段,但没有为创建元认知响应式 AI 系统提供全面指导。这样的系统应该包括哪些具体功能?它们应该如何检测不同的元认知模式?哪些支持策略对表现出不同模式的用户最有效?随着用户发展,支持应该如何随时间适应?这些设计问题在现有文献中基本上仍未得到回答。 这三个缺口共同表明,问题不在于简单地应用现有理论或实施已知解决方案,而是需要通过实证调查和人工制品创建来系统化地生成新的设计知识。问题表现出中等成熟度(新兴认识但理解不完整),而解决方案空间表现出低成熟度(很少或没有现有实施)。这一定位要求采用设计科学方法,既能推进我们对 AI 使用中元认知模式的理解(Ω 知识),又能提升我们设计支持元认知发展的系统的能力(Λ 知识)。我们的研究通过对跨越不同情境的 49 名用户进行系统化实证调查来解决这些缺口,使我们能够消费现有行为知识,同时以模式框架、设计需求和已实施系统架构的形式生产新的设计知识。

上述三个核心缺口在现有文献中具有多重表现形式。通过对元认知理论、信任框架与人工智能采用研究三条文献主线的系统梳理，可以更细致地识别既有研究在以下六个方面的不足，这些不足进一步支撑了本研究采用设计科学方法、聚焦元认知使用模式的必要性。

（1） 人工智能教育研究中元认知关注的系统性不足

多项元综述一致指出，人工智能教育研究对元认知的关注匮乏。Chiu 等（2024）在对 2014—2023 年间 126 篇关于教育中人工智能的综述进行综述时指出，尽管元认知常被提及为潜在收益，但系统性检视元认知过程的研究仅占少数。Nature 期刊在 2025 年对 57 项研究的元分析显示，对大学生的生成式人工智能影响“对元认知无显著作用”，其原因并非人工智能无法影响元认知，而是几乎没有研究对元认知进行测量（Rossi & Ferri，2025）。尽管该分析证实学业成绩（$g^{+}=0.633$）与高阶思维（$g^{+}=0.580$）具有较大效应值，但元认知维度的“无效应”实为“无测量”。针对人工智能赋能自我调节学习的定性系统综述仅纳入 14 项符合严格标准的研究，并指出研究者与用户均“倾向于优先采用快速且貌似最优的解决方案，而非较慢但更为可行的方案”，这种取向可能“削弱有效自我调节学习所必需的更深层认知与元认知过程”（Xu 等，2024）。

2024 年 ACM CHI 会议论文《生成式人工智能的元认知需求与机会》（The metacognitive demands and opportunities of generative AI）（Chilton 等，2024）明确提出，元认知“为理解与设计应对可用性挑战提供了宝贵视角”，同时指出此前研究中这一视角明显缺位。作者记录到多种元认知失灵，如用户即便在示范后仍回避有效提示设计，反映出调整人工智能心智模型的“元认知灵活性”不足。多项系统证据汇合，提示在人工智能教育研究中，关注元认知过程的研究可能仅占 7%至 10%，且多将元认知置于边缘结果，而非作为中心中介过程。

（2） 以结果为中心而非以过程为中心的研究主导

Vaccaro 等（2024）发表于《Nature Human Behaviour》的元分析整合了 106 项实验与 370 个效应量，集中体现了领域内以结果为中心的取向。研究普遍以准确率等性能指标量化人机协同是否优于人或机单独完成。几乎没有研究系统检视用户如何与人工智能共同决策，而仅问“结果是否更好”。作者建议拓展准确率之外的指标，如完成时间、成本与错误类型，但这些建议仍主要停留在结果层面，而非过程层面。其他多项系统综述亦证实这一模式。有关人工智能决策支持的研究强调效率与准确性（Deeva 等，2021）。教育心理领域指出，“学习分析中机器学习的早期应用主要聚焦于优化准确率与模型性能”，并承认“如果预测模型是不透明的‘黑箱’，即便准确，也难以提供可行动的洞见”（Ifenthaler & Yau，2020）。《Smart Learning Environments》亦批评了“黑箱取向”（Ifenthaler，2024），强调理解预测机制与预测准确同样重要，但实践上研究仍以改进预测为主，而非理解用户过程。测量实践同样反映这一偏向：研

究多采用前后测调查、日志行为与结果评估，鲜少使用思维口述、元认知访谈或实时监测来捕捉人工智能介入任务中的“思之所至”。

（3） 传统分类为何难以解释差异

信任与采用的文献共同揭示，以专家与新手、领域知识高低、或 STEM 与非 STEM 的分类，难以预测人工智能使用成效。《Nature Human Behaviour》的元分析发现，关键在于“相对性能”，而非专业水平。当人类在特定任务上优于人工智能时，人机团队更可能实现协同效应；当人工智能优于人类时，常见显著性能损失（Vaccaro 等，2024）。这表明，认识并校准“自我与人工智能的相对能力”的元认知过程，较静态的专业水平更能解释结果。2025 年的一项研究发现，传统的数字能力对人机协作结果无主效应，仅存在复杂交互，说明“熟练使用计算机者未必具备有效应用人工智能的技能”（Park 等，2025）。关于技能退化的认知研究进一步挑战专家假设：即便训练有素的专家也可能在常态性人工智能辅助下丧失任务技能，而受训者若长期依赖人工智能辅助则难以建立能力，专业性开始依赖人工智能而非稳定存在（Macnamara 等，2024）。

算法厌恶文献揭示非线性关系：高专长者与低专长者均可能表现出更强的算法厌恶，而中等经验者最可能采用人工智能并从中受益，呈倒 U 型模式，传统的专长二分难以容纳（Dietvorst 等，2018）。人口统计学的预测力更弱。已有研究记录人工智能算法会放大偏见，如低估少数族裔学生成功概率，或在死亡率预测中对少数群体出错，但这些准确性问题并不能告诉我们具体学生如何使用人工智能，或何种使用策略有效（Obermeyer 等，2019）。在精细考察个体差异的研究中，人工智能素养、提示工程与元认知监控的组内差异往往超过组间差异。首代大学生在恰当使用情境的自信较低，即便学术能力相当。心理特质如信任倾向、需要认知与自我效能可在专业与人口统计之外解释高达 24%的依赖行为方差（Schemmer 等，2024）。

（4） 采用元认知框架的理论论证

为何元认知可能在解释有效人机协作方面具有更强的解释力？四条理论论证可相互印证。第一，来自人机交互的分布式认知范式强调在真实工作情境中刻画计算介导的认知活动，而非孤立的个体变量，元认知过程框架正贴合这一诉求（Hollan 等，2000）。第二，来自人机协作的证据表明结果取决于对相对能力的识别与校准，提示元认知调校是关键中介。学习者必须监控自身理解、评估人工智能输出质量，并依据任务作出依赖调节，这些均为元认知过程。第三，来自学习科学的新近框架区分人工智能对学习的不同作用路径，如反转、替代、增强或重构，但要辨识何类学习者在何

种路径上受益，需要考察其在协作中的元认知参与，而非仅观察结果（Puentedura，2006）。第四，来自意义建构理论的研究指出，个体关于人工智能能力与边界的认知，将通过集体解释与信任建构影响组织层面的采纳，元认知意义建构在其中发挥关键作用（Weick，1995）。

综上，需要一种以用户认知过程为中心、能够刻画动态人机互动、纳入元认知意识与校准、并通过归纳方法发现自然发生模式、超越静态分类转向行为画像的框架。元认知理论在这些要求上具有独特优势：其提供了经验证的“监控—控制”构念，强调知识与行动之间的“知—行”缺口，允许在领域知识之外刻画个体差异，并与学习结果文献相衔接，证明元认知能力在智力之外预测成就。

（5）方法论启示：扎根理论的适配性

鉴于“学习者如何在与人工智能互动中开展元认知调节”这一问题具有探索性，且既有框架不足、使用模式在传统分类之外呈现高度异质，扎根理论为本研究提供了恰当的方法论路径。该方法已在多个人机交互议题中得到应用，能够通过系统编码自下而上地发现行为模式。其典型流程包括开放式编码（识别使用数据中的元认知行为）、主轴式编码（发展元认知策略类别）与选择式编码（构建元认知使用模式的理论模型）。近期研究利用扎根理论识别 XAI 系统中的解释对话要素（Ehsan 等，2019），发掘人机互动中的社会情感属性，包括信任与共情（Kim 等，2021），甚至尝试在扎根理论分析中引入人工智能辅助（Xiao 等，2023），显示该方法的演进与生命力。该方法尤其适合元认知研究，因为其天然面向探索、聚焦行为、重视过程与强调迭代。数据应包括使用过程中的思维口述、屏幕录制与回顾性访谈、交互日志、以及用于三角验证的元认知意识量表与绩效结果。分析宜围绕过程性问题展开：学习者如何在自用与求助人工智能之间作出选择？跨任务类型会涌现何种元认知策略？元认知意识如何随经验演化？在专业程度之外，哪些个体差异预测元认知有效性？

（6） 本研究的贡献定位

鉴于上述缺口，本文通过扎根理论方法，考察学习者在学习任务中的人机协作所呈现的元认知使用模式。不同于依据专业或人口统计类别，先行划分用户再比较结果的研究取向，本文直接观察不同经验水平的学习者“如何在使用人工智能的同时思考其思维”，即其采用了哪些元认知监控策略、如何将人工智能输出与学习目标相对照、如何跨任务进行依赖调节以及这些过程如何关联到学习成效。研究明确转向过程层面，关注中介学习结果的认知机制。研究亦放弃难以解释差异的专业分类，转而基于实际

的元认知策略，发掘行为模式。研究进一步将元认知理论、信任校准框架与人工智能采用研究三条相对独立的文献线索联结起来，强调元认知是连接个体差异、信任动态与使用效能的关键过程。

方法部分将详述数据采集与分析程序。研究发现将呈现通过系统编码得出的元认知使用模式。讨论部分将阐释这些模式如何解释传统专业分类无法解释的差异，如何为支撑元认知调节的教育人工智能工具设计提供启示，并在理论上将元认知定位为理解学习情境下人机协作的关键缺失环节。通过证明元认知能力较领域专长更能预测人工智能使用效能，本研究为教育人工智能研究开辟新方向，提示以元认知为导向的教学干预，并为“并非所有学生都能等量受益于人工智能”的直觉提供理论支撑，原因在于他们在与这一强大而复杂的工具互动时的元认知参与存在根本差异。

3.3 研究方法

3.3.1 设计科学研究方法论

本研究遵循 Hevner 等人（2004）提出的设计科学研究范式，在 Peffers 等人（2007）的六阶段 DSR 过程模型中，当前论文聚焦于“定义解决方案目标”阶段。 我们通过系统化的实证调查来回答：元认知响应式 AI 系统应该实现什么功能？ 需要识别和响应哪些用户模式？如何将实证发现转化为可操作的设计需求？ 为确保研究的严谨性，我们在方法设计中明确遵循 DSR 的核心指南：（1）设计为人工制品：生产模式框架、设计需求、评估方法等多层次人工制品；（2）问题相关性：解决 AI 教育中的元认知支持缺失问题；（3）设计评估：采用描述性评估（本文）与实验/观察评估（后续论文）相结合；（4）研究严谨：运用系统化编码、多重效度检验、定量补充分析；（5）研究贡献：生产Ω知识（模式理论）与Λ知识（设计需求）。基于上述DSR定位，我们采用解释主义取向的定性研究设计（Walsham, 1995）， 以扎根理论方法识别元认知使用模式，并通过反向链结法推导设计需求。以下各小节详述具体实施程序。

3.3.1 受访者招募与样本选择

（1）抽样策略

本研究采用目的性、最大差异抽样（Patton，2002），以覆盖广泛的人工智能使用情境、能力水平与学科领域。该策略契合我们的研究目标，即识别能够跨越传统用户分类（专家与新手、理工与非理工）的稳健模式。我们重点在五个维度上寻求差异：①受教育程度（本科至博士）；②职业角色（学生、研究人员、行业从业者、创业

者）；③学科背景（理工、商科、人文、创意类）；④人工智能使用频率（每日高频至偶尔使用）；⑤地理与文化情境（以新加坡与中国为主，辅以部分国际样本）。

招募经由三条渠道开展。其一，通过厦门大学与南洋理工大学的院校发布招募广告，明确寻求每周至少 5 次将 ChatGPT、Claude、Gemini 等生成式人工智能用于学术或专业工作的参与者。其二，采用滚雪球抽样（Biernacki 与 Waldorf，1981），请初始受访者转介不同使用模式或职业背景的同事。其三，在面向人工智能工具的线上社群发布招募启事（Reddit r/ChatGPT、LinkedIn 专业群组），以覆盖学术样本中相对不足的在职群体。为避免技术熟练度造成的系统性偏差，我们刻意不集中于计算机科学或人工智能相关领域。

纳入标准从宽以提高生态效度：受访者需满足以下条件。①年满 21 周岁；②将生成式人工智能用于学习或知识工作而非纯娱乐；③愿意以 45–75 分钟时长开放讨论其使用实践；④能够以英语或普通话交流。我们不设人工智能使用能力门槛，预期新手的困难与专家的策略同样具有信息价值。排除标准仅限于两类人群：仅将人工智能用于娱乐用途者，以及人工智能企业员工，以避免营销话语影响。

（2） 最终样本特征

最终样本包括 49 名受访者（学生 23 人，职场人士 26 人），访谈时间为 2025 年 4 月至 2025 年 8 月。表 3-1 展示了样本的人口学特征。

表 3-1: 受访者人口学特征（N＝49）

指标	分布	备注
教育程度		
博士研究生	9 人（18%）	市场营销、计算机科学、金融、工程、光学、材料
硕士研究生	7 人（14%）	数据分析、数据科学、医疗、文科、工商管理
本科生	6 人（12%）	计算机、电子商务、管理科学、电气电子
职场人士（毕业后）	26 人（53%）	博士后与研究员 6 人（16%），行业从业者 20 人（45%）
年龄分布		
21–25 岁	16 人（33%）	主要为本科与硕士学生
26–30 岁	20 人（41%）	多为博士研究生与初入职场者
31–40 岁	11 人（22%）	职业中期的研究员或管理岗位
41–45 岁	2 人（4%）	资深从业者
性别		
女	22 人（45%）	
男	27 人（55%）	
学科背景		
理工	26 人（53%）	计算机、工程、数据科学、化学、物理、生物医学

商科与管理	17 人（35%）	会计、金融、市场、电子商务
人文与教育	3 人（6%）	教育、学习科学、人文学科
创意与其他	3 人（6%）	广告、设计、健身与内容创作
主要使用的人工智能工具		
仅使用 ChatGPT	31 人（63%）	最常见单一工具
多工具组合	18 人（37%）	ChatGPT 搭配 Gemini、Claude、DeepSeek 等
使用频率		
每日（每周 5–7 天）	37 人（76%）	
每周数次（3–4 天）	10 人（20%）	
每周 1-2 次	2 人（4%）	
地域情境		
新加坡	24 人（49%）	南洋理工大学及社会从业者
中国内地	15 人（31%）	厦门大学、清华大学、中国人民大学等及从业者
其他国际	10 人（20%）	亚洲地区外籍人士

样本规模依据理论饱和确定（Glaser 与 Strauss，1967）。在每周组会中持续评估新出现的行为模式是否仍在增加。当访谈进行至约 35 场时，初步模式出现稳定迹象，后续 14 场用于验证模式的普遍性与边界条件（例如极低频使用者、量化交易等高度专业领域）。此样本规模与定性信息系统研究识别模式的常规建议相符：Creswell（1998）建议现象学研究 20–30 例，而 Guest 等（2006）指出同质样本 12–15 例即可达到饱和，异质样本通常需要 30 例以上。

3.3.2 数据收集程序

（1） 访谈提纲的制定

我们依据元认知理论（Flavell，1979；Schraw 与 Dennison，1994）并参考人机互动研究（Bødker，1996；Dourish，2004），设计了半结构化访谈提纲。提纲经三轮试访（5 名参与者，未纳入正式分析）以优化问题表述、避免诱导与控制时长。正式提纲包括四个部分：

第一部分：人工智能使用脉络化（5–10 分钟）。以开放式问题建立关系并了解受访者的人工智能使用全景，例如“请回顾您最初开始使用 ChatGPT 等工具的情境”与“请描述您典型的一天何时使用人工智能、何时不使用”。此部分兼顾建立信任与识别多样化使用情境，以便后续追问情境依赖的策略差异。

第二部分：具体事件的细致复原（25–40 分钟）。这是数据收集的核心。我们请受访者选取两段近期交互事件，一段自评成功、一段自评具有挑战性，并进行细节复原。借鉴关键事件技术（Flanagan，1954）与回顾性口语报告法（Ericsson 与 Simon，1993），引导其逐步陈述输入内容、当时思考、预期、对输出的判断以及后续动作。此过程有助于呈现通常自动化的元认知过程，包括规划、监控、评估与调节。当受访者仅给出概述性描述时（如“我请 GPT 做文献综述”），我们使用追问以获得认知细节，例如“在提问前是否有既定使用规划”、“如何决定输出是否足够好”、“是否进行了核查”、“事后会否做不同选择”。这些追问旨在触达监控与评估过程。

第三部分：策略与模式的自我陈述（10–20 分钟）。在具体案例之后，引导受访者升维概括其策略，例如“您对何时使用人工智能与何时自行完成是否有规则”、“您的使用方式是否随时间发生变化，如何变化”、“有哪些任务您后来不再使用人工智能”。同时讨论信任动态，例如“在 0–100% 的刻度上，您对人工智能输出的信任度是多少，是否随情境而变”。此部分主要获取受访者的显性元认知知识（Schraw 与 Dennison，1994）。

第四部分：困难与理想系统设计（5–15 分钟）。以反思性问题收束，例如“在学习中使用人工智能最令您困扰之处”、“您希望人工智能具备而目前尚未具备的能力”、“若可为学习目的重新设计人工智能系统，您会改变哪些方面”。此部分直接从痛点中生成初步设计需求。

考虑到地域分布，受访均通过 Zoom 在线进行。线上访谈具有方法优势：受访者可便捷共享屏幕展示工作流，录制资料在征得同意的前提下同时保留了语言与视觉信息。访谈以受访者偏好的语言进行（英语 27 例，普通话 22 例），由双语研究者主持，确保语言舒适度与表述真实。中文访谈后续转录并译为英文以便统一编码，翻译准确性通过对 15% 语料的回译抽查完成。

（2） 访谈实施与研究者反思

三名受过训练的研究者（本研究的第一作者、第二作者与一名科研助理）承担访谈工作，具有学习科学与人机交互背景。数据收集前，研究团队开展反思练习（Malterud，2001），以呈现关于“有效使用”的先入之见。我们承认最初假设包括专业能力可能预测使用质量、理工背景用户更具策略性等，这些假设在数据中受到挑战。访谈过程中采取非评判立场，明确表示关注多样实践而非能力评估。对于看似次优的行为（如不加核查地接纳输出），访谈者以探究性态度追问其理由与感受。

单场访谈时长 45–93 分钟，平均 56.2 分钟，标准差 12.7 分钟。时长差异与受访者的策略复杂度与表达详尽度相关。所有受访者依据机构规范获得 50 新元现金作为酬谢。

（3） 数据存证

每场访谈形成三类资料：①音频记录（ 49 人都同意录音）；②经专业转写并由研究者复核的逐字稿；③访谈者备忘录，即刻记录情境观察、初步主题与互动反思。备忘录在分析阶段用于辨析语气、情绪与意涵，有助于弥补文本的不足。

全部逐字稿合计约 743,291 字，单场平均约 15,168 字。中文逐字稿由持证双语译者翻译为英文，研究团队以概念等值为原则进行复核。例如将“认知外包”译作“cognitive outsourcing”，以保留理论含义。在 3 例中，受访者于访谈中出现中英夹杂，我们在逐字稿中保留原文并以方括号提供英文翻译。

3.3.3 数据分析：迭代编码与模式识别

（1） 分析框架与编码结构

分析遵循建构主义扎根理论（Charmaz，2014），在理论敏感化的前提下进行归纳。我们结合元认知框架（Flavell，1979；Azevedo 与 Hadwin，2005；Winne 与 Hadwin，1998），开展“三阶段”分析流程：开放编码生成初始范畴，聚焦编码构建元认知过程分类体系，理论编码识别更高层次的使用模式。此一迭代过程既保持材料扎根，又借助理论概念解释不可直接观察的认知现象。

阶段一：开放编码与初步归类。在数据收集并行开展常比法（Glaser 与 Strauss，1967）。每场访谈后撰写分析备忘录（Saldaña，2021），记录初步主题、意外发现与与既有个案的联系。前 10 份逐字稿完成后，研究团队在 NVivo 15 中进行逐行开放编码，生成 387 个初始代码，覆盖行为、认知与态度等现象。例如“手工任务分解”“外部交叉核查”“情境化的信任表达”“技能退化担忧”“在高质量建议下仍拒绝采用”等。为降低复杂度，我们将初始代码汇总为 12 个临时类别，涵盖任务取径、核查行为、信任表达、学习取向、错误响应、工具选择、时间管理、技能自觉、协作隐喻、伦理考量、情感反应与情境差异。该归类并非预设，而是在元认知“规划、监控、评估、调节”的敏感化框架下自然形成。

阶段二：聚焦编码与元认知过程分类。考虑到 387 个代码过于细碎，不利于模式识别，我们开展聚焦编码（Charmaz，2014），据此凝练为具理论意义的范畴。参考 Flavell（1979）与 Schraw 与 Dennison（1994），构建了适用于人机协作情境的元认

知过程分类体系（见图 3-1）。该体系包含四类高阶元认知过程与十二个具体子过程，涵盖用户在与生成式人工智能协作时所展现的认知调节活动。

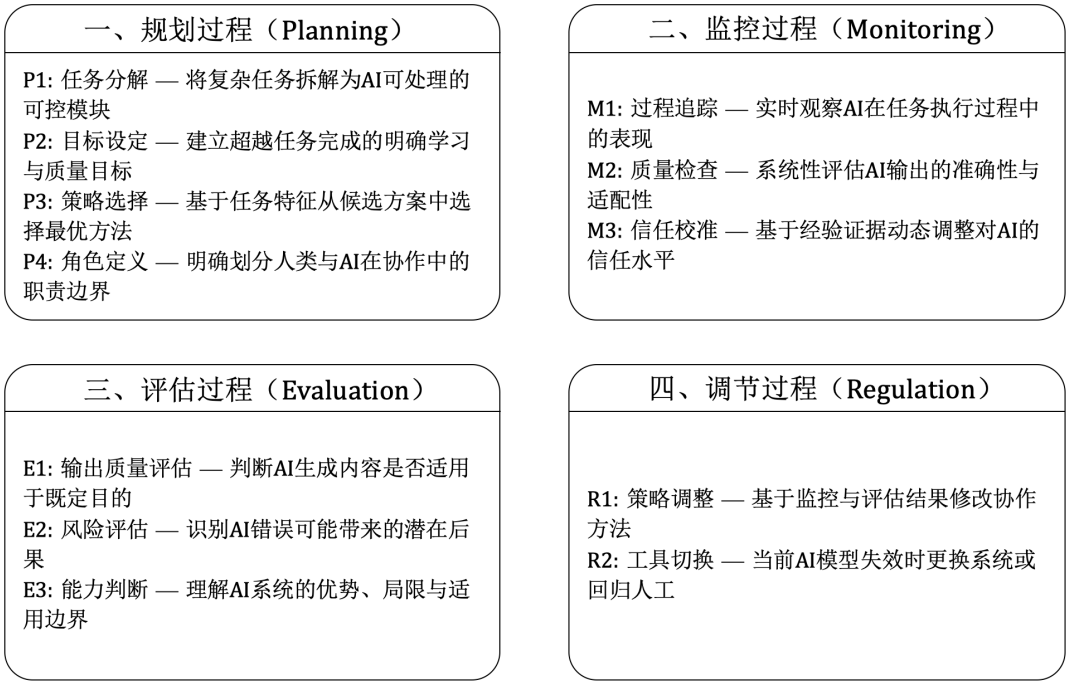

图 3-1 人机协作中的元认知过程分类框架

资料来源：改编自 Flavell, 1979; Azevedo & Hadwin, 2005

为保证编码的系统性与可复制性，对每一位受访者，我们制定了三级证据强度判定标准（见表 3-2）。该标准综合考虑了陈述的明确性、行为证据的一致性以及跨情境的重复性，确保编码结果既植根于数据又具有理论一致性。具体为：✓✓✓ 表示强证据，✓✓ 表示中等证据，✓ 表示弱证据。例如“我不会整单交给 GPT。我会先把一万字压缩到两千字，再分节输入”的表述，对 P1 与 P4 记为强证据。两名研究者独立完成 49 份逐字稿的编码，并以 Cohen κ 评估一致性。初始 κ＝0.71，达实质一致。分歧主要出现在中等与弱证据的区分上，经回查逐字稿与备忘录讨论后达成一致。最终形成 49×12 的元认知子过程矩阵，并计算每位受访者的元认知复杂度得分（0–12），即具有中等及以上证据的子过程计数之和。

表 3-2：元认知过程证据强度编码标准

符号	强度	判断标准	示例
✓✓✓	强	• 受访者多次明确陈述该过程 • 提供一致的行为证据 • 在至少 2 个不同情境中出现 • 包含过程本身的元陈述	“我从不把整个任务交给 GPT”（多次强调）+ 描述了详细的分段流程 +

			在学术写作和代码开发中均使用
✓✓	中	• 清晰的行为描述（至少一次） • 或单次明确陈述 • 有具体实例支撑	“我会逐段检查 AI 的修改，对比原文和改后版本”（描述了具体操作流程）
✓	弱	• 从上下文推断 • 或简短提及 • 缺乏详细说明	通过“后来我发现了错误”间接推断进行了某种形式的验证，但未明确描述验证过程

阶段三：聚类与诠释结合的模式识别。 在获得元认知画像后，我们采用数据驱动的聚类分析与理论驱动的质性诠释相结合的双轨方法识别使用模式。首先对 49×12 矩阵（49 名参与者×12 个元认知子过程的评分）进行层次聚类分析（Ward 法，欧氏距离）。我们系统地评估了 3 至 7 类解的质量。树状图（dendrogram）分析显示多个潜在的切割点，表明数据具有层次结构。为量化不同聚类数的质量，我们计算了轮廓系数（silhouette coefficient），该指标综合衡量类内凝聚度和类间分离度（Rousseeuw, 1987）。3 类解的轮廓系数为 0.48，4 类解为 0.51，5 类解为 0.49，6 类解为 0.52（最高），7 类解为 0.47。6 类解的轮廓系数最高，虽属“合理结构”（Kaufman & Rousseeuw, 1990：0.51-0.70）而非“强结构”（0.71-1.0），但这一数值恰恰反映了元认知使用模式的固有复杂性：61%的参与者展现混合模式特征，同一个体在不同情境下可能切换模式，元认知行为本质上是连续谱而非离散类别。随后依据 Miles 等（2014）关于混合方法的建议，我们不以算法聚类为唯一依据，而是将其视为需要深度质性验证的初步假设。研究团队（主要研究者、1 位教育心理学协作者、1 位资深信息系统研究者）经六周迭代（12 次研讨会议），反复阅读各类群的逐字稿，检验其是否具备清晰的行为签名与内在逻辑，并能否对应现实中的用户类型或情境性策略。对于 3 个边界案例（算法分类与质性判断不一致），团队通过讨论达成共识并进行人工调整。综合聚类分析和质性验证的双重证据，最终归纳出六类模式：模式 A：战略性分解与控制（18 人，37%）；模式 B：迭代优化与校准（4 人，8%）；模式 C：情境敏感的适配（16 人，33%）；模式 D：深度核验与批判性介入（4 人，8%）；模式 E：教学化反思与自我监控（7 人，14%）；模式 F：无效与被动使用。值得说明的是，模式 F 虽在质性分析和教师访谈中被识别为一种存在的使用模式，但在本研究的直接观察样本（N=49）中未有受访者将其作为主要使用策略。该模式的识别主要基于：（1）部分受访者描述的"

早期尝试阶段"的行为特征，以及（2）教师对学生 AI 使用中观察到的问题性模式。模式 F 在直接观察样本中的缺失可能反映了样本选择效应（愿意参与深度访谈的用户通常已发展出一定的元认知策略），但其作为理论对比案例对于界定有效使用的必要条件具有重要意义。因此，后续的定量补充分析聚焦于前五种主要模式（A-E）。我们为各模式界定必要与充分条件（见第 3.4.2 节），以便在承认混合模式存在的前提下进行系统分类。对于呈现混合模式的受访者（30 人，占 61%），以高风险任务中的主导模式作为分类依据（因高风险情境更能揭示深层元认知倾向），并记录次级模式及情境触发因素。

（2） 信效度保障

为提升研究的严谨性与可信度（Lincoln 与 Guba，1985），我们采取四项策略：（1）成员核对。将初步模式描述与分类发送给覆盖不同模式的 10 名受访者，请其反馈“是否符合自我实践”。9 人认可，1 人提出在不同情境下表现为模式 A 与 C 两类。基于此，我们系统编码情境驱动的模式切换，发现 41% 用户存在此类行为。成员核对在访谈后 1 个月进行。（2）专家审阅。邀请三位外部学习科学学者（教育心理学两位、认知科学一位）评阅 12 子过程分类与模式框架的一致性、逻辑性与证据充分性。专家肯定其理论一致性，并建议强调 F 类并非“策略”，而是“策略缺位”，据此我们调整了概念表述。（3）反例分析。主动搜索与既定框架相冲突的证据（Patton，2002）。例如我们最初假设博士生将集中于模式 A 或 模式 D，但发现存在博士生呈现模式 F 的行为。我们据此修正理论主张：专业能力并不预测模式归属。（4）审计轨迹。完整保存编码手册迭代版本（7 版）、备忘录（4 个月内 143 份）、团队讨论纪要（18 次）与模式定义修订的决策记录，以确保方法透明与可复核（Morse，2015）。

3.3.4 定量补充分析

在完成三阶段质性编码后，本研究实施了系统化的量化补充分析，以增强研究发现的可验证性与可推广性。这一混合方法策略遵循 Maxwell（2010）提出的整合性研究设计原则：量化分析用于补强而非取代诠释性分析，通过统计推断为模式流行率、跨模式差异和预测因素提供实证证据，同时保持对质性数据丰富性和情境复杂性的尊重。本小节详细说明质性数据的量化转换流程、统计分析方法及其理论与方法论依据。

（1）质性数据的系统化量化转换

编码矩阵的构建与操作化定义。基于前文描述的 12 子过程元认知分类体系（规划 4 项、监控 3 项、评估 3 项、调节 2 项），我们构建了 49×12 的编码矩阵作为质性-

量化转换的基础架构。矩阵的每个单元格记录特定参与者在特定元认知子过程上的证据强度，采用四级序数量表进行系统性编码：

强证据（✓✓✓，赋值 3 分）：访谈中出现多次明确陈述，配以详细行为描述，且跨情境展现一致性。例如，受访者 1 关于任务分解的陈述："我从不把整个任务交给 GPT。我先手动将万字文本压缩到 2000 字，然后逐节喂给 GPT 进行润色。我需要逐段比对编辑，因为 AI 在处理大量内容时可能会偷工减料。"这段叙述包含明确的策略陈述、具体的工作流程描述（40 分钟的手动压缩过程）以及背后的认知理由，满足强证据的全部标准。

中等证据（✓✓，赋值 2 分）：访谈中出现清晰的行为描述或单次但明确的策略陈述。例如，某参与者提及"需要保持对核心论点的控制"作为使用 AI 的指导原则，虽仅陈述一次但表述明确，且在后续讨论中行为表现与该原则一致。

弱证据（✓，赋值 1 分）：需要从上下文推断的行为，或仅有简短、模糊的提及。例如，参与者在描述工作流程时顺带提到"也会看看别的方案"，但未详细说明策略选择的标准或过程。

无证据（赋值 0 分）：访谈中完全未出现该子过程的相关行为或认知陈述。

两位独立编码员对全部 49 份访谈进行编码，通过对 10%样本（n=5）的双重编码计算编码者间信度：Cohen's Kappa = 0.71（$p < .001$），达到 Landis and Koch (1977) 建议的“实质性一致”水平。分歧主要集中在区分中等证据（✓✓）与弱证据（✓）时，解决方式为两位编码员共同回顾原始访谈录音，参考访谈者备忘录以获取情境信息，通过讨论达成共识。最终所有 588 个编码单元（49 参与者×12 子过程）均实现一致性评定。

复合指标的构建与计算规则。从编码矩阵派生三类关键量化指标，每类指标均具有明确的理论依据和操作化规则：

1）元认知复杂度分数（Metacognitive Complexity Score, MCS）

该指标测量参与者跨元认知子过程维度的参与广度，反映 Flavell (1979)提出的元认知多维性特征。计算公式为：

$$MCS = \frac{\sum_{i=1}^{12} I(\text{Score}_i \geq 2)}{12}$$

其中，$I(\cdot)$为指示函数，仅当子过程 i 的证据强度达到中等（✓✓）或强（✓✓✓）水平时取值为 1。弱证据（✓）不计入得分，因为其可能反映偶然行为而非稳定的元认知策略（Schraw 与 Dennison, 1994）。得分范围为 0-12，其中 0 表示无元认知证据，12 表示全部子过程均展现中等以上参与强度。

2）验证强度指数（Verification Intensity Index, VII）

该指标测量参与者在AI输出验证方面的系统性和彻底性，对应Lee和See(2004)信任校准框架中的“监测行为”维度。基于访谈数据中识别出的验证行为类型，我们构建了 10 项检查清单，每项采用 0-1 连续评分：

[1] 主动追踪 AI 输出变化（如要求 track changes 功能）

[2] 接受前验证事实准确性（0=从不验证，0.5=选择性验证，1=总是验证）

[3] 跨来源交叉检查（如"三角验证"：AI→测试→在线资源）

[4] 测试生成的代码或公式（若适用）

[5] 验证引用文献的存在性（若适用）

[6] 多模型比较（0=从不，0.5=偶尔，1=系统性）

[7] 高风险任务中咨询领域专家

[8] 检查输出的内部逻辑一致性

[9] 评估回答的合理性/常识符合度

[10] 对 AI 保持系统性怀疑态度

两位研究者根据访谈中的具体行为描述独立评分，讨论至达成一致。对于不适用的项目（如某参与者从不使用 AI 处理代码任务，则[4]项标记为N/A），采用比例调整公式：$VII = (\text{实际得分}/\text{适用项目数}) \times 10$。此公式确保不同任务类型的用户可比性。

3）行为频率与变异性指标

为捕捉模式特征的其他关键维度，我们从访谈叙事中提取三类行为频率指标：（1） 任务分解频率：参与者描述分解行为的任务数量与访谈中讨论的总任务数量之比，操作化定义为至少明确提及将任务拆分为2个以上子任务的情况数。（2） 信任变异系数：参与者跨不同情境（学术高风险、常规工作、个人低风险、创意探索）报告的信任评分（0-100%量表）的标准差，用于量化模式 C 的“情境依赖性信任校准”特征。（3） 独立工作比例：参与者报告独立完成的任务占其近期 AI 相关任务总数的百分比，基于自我报告与行为描述的三角验证。该指标支撑 MR16（技能退化预防系统）的实证基础。

（2）统计推断方法的选择与应用

描述性统计与分布检验。对所有连续变量（MCS、VII、年龄等）计算均值、标准差、中位数、四分位距和偏度/峰度系数。Shapiro-Wilk 正态性检验显示 MCS 在 A-模式 E 中近似正态分布（W = 0.96, p = .18），但 Pattern F 显著偏离（W = 0.81, p = .04）。因此，涉及 Pattern F 的比较采用非参数检验（Mann-Whitney U 或 Kruskal-Wallis H），其余采用参数检验（t 检验或 ANOVA）。

模式流行率的统计推断。为验证五种主要元认知模式的分布是否偏离随机期望，我们进行卡方拟合优度检验（chi-square goodness-of-fit test）。零假设为五种模式均匀分布（每种模式占 20%），观察频数为 A=18(37%), B=4(8%), C=16(33%), D=4(8%), E=7(14%)。

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = \frac{(18-9.8)^2}{9.8} + \frac{(4-9.8)^2}{9.8} + \frac{(16-9.8)^2}{9.8} + \frac{(4-9.8)^2}{9.8} + \frac{(7-9.8)^2}{9.8}$$

计算得$\chi^2(4) = 18.44, p = .001$，拒绝均匀分布假设。进一步的二项精确检验（binomial exact test）验证模式 A 是否显著超出"主要模式之一"的预期（$H_0: P(\text{Pattern A}) = 0.20$），结果为 p=.005（双侧），确证模式 A 为最普遍有效模式。

跨模式差异的方差分析。

单因素方差分析（one-way ANOVA）检验五种主要模式在元认知复杂度得分（MCS）上的均值差异。方差齐性检验（Levene's test）显示$\text{F}(4,44) = 6.23, \text{p} = .016$，违反方差齐性假设，因此采用 Welch's ANOVA 作为稳健替代：

$$F_{\text{Welch}}\,(4, 24.3) = 42.73, p < .001, \omega^2 = 0.81$$

效应量ω^2（omega squared）选择而非常用的η^2（eta squared），因为前者在样本量不等且方差异质时提供更准确的总体效应估计（Olejnik & Algina, 2003）。ω^2=0.81 表示模式归属解释了 81%的 MCS 变异，属于极大效应（Cohen, 1988）。事后多重比较采用 Games-Howell 程序（适用于方差不齐情况），仅报告理论关键对比以控制家族误差率（familywise error rate）：

各有效模式（A, B, C, D, E）均展现显著的元认知觉察水平（MCS ≥ 5）。高元认知模式组（A, D, E）的平均 MCS 得分显著高于中等元认知模式组（B, C），但组内模式之间无显著差异：

模式 E 对比模式 A：$\Delta M = 0.6, 95\%\ \text{CI}\ [-0.3, 1.5], p. 34\ (\text{ns})$

模式 E 对比模式 D：$\Delta M = 0.2, 95\%\ \text{CI}\ [-0.9, 1.3], p = .67\ (\text{ns})$

模式 B 对比模式 C：$\Delta M = -0.4, 95\%\ \text{CI}\ [-1.6, 0.8], p = .58 (\text{ns})$

置信区间报告提供效应大小的精度估计，比单一 p 值提供更丰富信息（Cumming, 2014）。

人口学变量与模式成员身份的关联检验。为检验传统用户分类（教育水平、学科背景、年龄）是否预测模式 A 成员身份，我们进行两类统计检验：

（1）教育水平（研究生对比本科）和学科背景（理工类对比非理工类）：卡方独立性检验

构建 2×2 列联表，期望频数通过边际分布计算：$E_{ij} = \left(n_{i\cdot} \times n_{\cdot j}\right)/\text{n}$

（2）年龄：点双列相关（point-biserial correlation）

年龄为连续变量，模式 A 成员身份为二分变量，采用 Pearson 相关的特殊形式：

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}}$$

其中M_1和M_0分别为模式 A 组和非模式 A 组的年龄均值，s_n为全样本年龄标准差

（3）多变量预测模型：二元 Logistic 回归

为探索元认知复杂度的潜在预测因子，我们对 49 位参与者进行了探索性回归分析。虽然样本量有限，但鉴于：(1) 本研究为质性主导的混合方法设计，定量分析作为质性发现的补充验证；(2) 观察到的大效应量($f^2 = 1.03$)使得 N=49 的样本具有充足的统计功效（100%）来检测主要效应；(3) 核心发现达到高度显著水平(p<.001)，表明结果稳健可靠；我们认为这一分析能够提供有价值的初步证据。我们将结果解读为“模式识别”而非“精确参数估计”，并建议未来研究使用更大样本验证这些发现。为同时控制多个潜在混杂因素并比较各变量的相对预测力，我们构建二元 Logistic 回归模型。因变量为元认知觉察水平，通过 MCS 中位数分割法二分化（高≥7 分对比低<7 分）以满足 Logistic 回归的二分因变量假设。虽然二分化会损失信息，但使结果更易解释且符合"高/低元认知觉察"的理论意义（Hosmer et al., 2013）。

模型设定为七个预测变量同时进入模型（enter method），避免逐步回归可能导致的过拟合和不稳定性（Harrell, 2015）：连续变量（均已中心化以减少多重共线性）：年龄（years）和 AI 使用频率（days per week）；二分类变量（效应编码）：教育水平（研究生=1，本科=0），学科背景（STEM=1，非 STEM=0），教学经验（有=1，无

=0），经历重大 AI 错误（是=1，否=0）和接受过元认知训练（是=1，否=0）。模型方程为：

$$\text{logit}\,(P(Y=1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_7 X_7$$

其中$Y=1$表示高元认知觉察。

3.3.5 设计需求的推导

在模式识别之外，我们以可落实的元认知支持系统设计需求为次级目标。采用反向链结方法（Walls 等，1992），从观察到的行为回推必要的系统赋能。对每一类模式，我们提出两个关键问题："哪些功能可使有效策略更容易被实施，或使其变得不再必要""哪些功能可预防无效行为的发生"。经由此过程生成 87 项初步设计构想，随后以亲和图法（Holtzblatt 等，2005）整合为 19 条元需求。我们依据四个标准进行优先级排序：用户影响（频率与严重度的乘积）、技术可行性、理论支撑强度以及对采纳的促进或阻滞作用。最终形成两条"关键"需求（MR13 透明化不确定性呈现；MR23 隐私保护型架构）、13 条"高优先级"与 6 条"中优先级"需求，详见第 4.3 节。

3.3.6 伦理考量

本研究在数据收集前获得机构伦理审查批准（IRB 编号：IRB-2025-446）。所有受访者在充分知情后以书面形式同意参与。核心伦理安排包括：（1）保密与去标识化。鉴于受访者可能涉及上级不认可的工作实践，我们实施严格的保密措施。所有受访者以编号（I1–I49）标识并贯穿全文。逐字稿进行去标识化处理，删除姓名、单位与可识别信息。涉及具体项目时改以通用描述替代。数据存储于加密并受限访问的服务器。（2）受访者福祉。部分受访者在访谈中表达对自身使用习惯的焦虑，尤其是 F 类倾向者因访谈而意识到过度依赖的可能。访谈者以共情回应，将研究兴趣定位于理解多样实践而非评判能力，并在访谈结束后提供一份有效使用策略的资源清单。（3）权力关系。考虑到学生可能出于学术评价压力而粉饰自我表现，我们在开场时明确非评估立场。针对初级研究人员讨论其与导师之间的限制时，引导聚焦个体经验而非机构批评。（4）数据留存与共享。音视频资料依照伦理要求保存 5 年后删除。去标识化逐字稿将长期归档以备后续分析。研究成果发表后计划在 OSF 提供去标识化数据集，以便二次分析与可复核。

3.3.7 局限与方法反思

本研究在方法上存在若干局限，需予以说明。（1）横截面取向。访谈捕捉的是单一时间点或近月的回顾性图景，无法确定模式在更长时期内的演化机制与转换触发

因素。未来可开展 6–12 个月的追踪研究。（2）自陈的有效性。尽管使用了口语报告等方法，受访者的回忆可能受记忆能力与社会赞许偏差影响。观察性方法（如实时屏幕录制与同步思维口述）有助于三角互证，但亦可能引入反应性效应。（3）样本构成。样本主要为新加坡与中国的研究生与职场人士，且以 ChatGPT 为主。研究结论在以下人群与情境中可能不具可迁移性：中小学学习者、不同教育理念的文化情境、领域专用系统使用者（如 GitHub Copilot）、以及未来可能改变人机动态的新能力。建议在多元情境中开展重复研究。（4）模式稳定性的假设。本研究将模式作为相对稳定的策略类型进行分析，然而有 41% 的受访者表现出情境驱动的模式切换。尽管我们对该现象进行了编码，本框架仍可能高估连贯性而低估情境流动。未来研究可将使用轨迹建模为动态过程。（5）研究者立场。团队成员的学科背景与价值取向偏向主动学习、元认知发展与人类能动性，这在一定程度上影响了对各模式的评价。我们通过反思努力降低偏差，但理论视角不可避免地塑造了诠释。来自产业导向的研究者可能对以效率为目标的行为给予更正面的评价。尽管存在上述局限，本研究通过扎实的方法设计与质量控制，包括清晰的理论奠基、多样化抽样、系统性编码、定量补充与多重效度策略，为关于生成式人工智能在以学习为导向的知识性任务中的使用模式提供了可信且可迁移的证据。

3.4 研究发现

基于对 2025 年 4 月至 2025 年 8 月期间开展的 49 场深度访谈的分析，本文识别出六类彼此区分的元认知使用模式。这些模式是由用户在与人工智能交互的任务过程中持续运用的元认知过程所构成的有机组合，涵盖规划、监控、评估与调节四个维度（Flavell，1979；Schraw 与 Dennison，1994）。分析结果呈现一项与技术采纳领域常识相悖的重要发现。用户身份并不预测使用成效。研究生既可能表现出高度有效的策略性使用（A 类），也可能出现问题性过度依赖（F 类），而本科生则可能展现与资深研究者相当的元认知精细化水平。此一分布挑战了人–AI 交互文献中的普遍假设，即专业水平能够可靠预测恰当的技术使用（Glikson 与 Woolley，2020；Amershi 等，2019）。相反，数据表明，有效的 AI 协作在本质上是元认知成就，而非专业等级属性。区分成功与失败用户的关键不在于其学科知识、受教育程度或职业资历，而在于其为维持学习、能动性与批判性思维而部署的具体元认知策略。本节依次呈现六类模式，

并进行跨模式比较以揭示区分有效与无效使用的元认知维度，最终提出面向“元认知协作型代理”（Metacognitive Collaborative Agents，MCAs）的十九条基于证据的设计需求。

3.4.1 受访者元认知特征与模式分布总览

在详细阐述各模式的质性特征之前，我们首先呈现所有受访者的元认知复杂度评分与模式归属综合分析结果（表 3-3）。此总览表基于前一节描述的系统化编码程序生成，为理解后续质性叙事提供量化参照坐标，并支持本研究的核心论点验证,即有效 AI 使用是元认知成就而非专业等级属性。

表 3-3：所有受访者的元认知复杂度评分与模式归属综合分析结果表

受访者编号	身份类别	职业 / 学习领域	性别	年龄	元认知复杂度	主要使用模式
I1	学生	市场营销博士生	女	26	高（8/12）	模式 A
I2	学生	会计学博士生	女	26	高（7/12）	模式 C
I3	学生	工商管理博士生	男	26	极高（9/12）	模式 A
I4	学生	会计学硕士生	女	21	高（7/12）	模式 C
I5	学生	数据分析硕士生	男	23	高（7/12）	模式 E
I6	学生	医疗数据科学硕士生	男	25	高（8/12）	模式 A
I7	学生	电子商务本科生	男	21	高（7/12）	模式 A
I8	学生	计算机科学本科生	女	21	高（8/12）	模式 D
I9	学生	管理科学本科生	女	21	高（7/12）	模式 B
I10	职业人士	金融学副教授	男	37	中（5/12）	模式 C
I11	职业人士	产品经理	女	28	中（5/12）	模式 C
I12	职业人士	金融风控专员	女	30	中（6/12）	模式 C
I13	学生	人工智能算法工程师	男	22	极高（10/12）	模式 A
I14	职业人士	市场分析师	女	21	中（6/12）	模式 A
I15	职业人士	工业自动化战略分析师	男	37	中（6/12）	模式 B
I16	学生	计算机科学博士生	男	24	极高（10/12）	模式 A
I17	学生	金融学博士生	男	23	高（8/12）	模式 D
I18	职业人士	航空航天研究员	女	30	高（8/12）	模式 C
I19	职业人士	项目经理	女	27	高（7/12）	模式 C
I20	职业人士	教育学研究员	女	30	高（8/12）	模式 A
I21	学生	人机交互硕士生	女	26	高（8/12）	模式 E
I22	学生	深度学习博士生	男	25	极高（10/12）	模式 A
I23	职业人士	Web3 交易平台运营	男	30	中（5/12）	模式 C
I24	职业人士	软件工程研究员	女	30	极高（9/12）	模式 E
I25	职业人士	人工智能与教育研究员	女	25	极高（10/12）	模式 E
I26	职业人士	电商创业者	女	27	高（7/12）	模式 C
I27	职业人士	蛋白质与化学研究者	女	23	中（6/12）	模式 C
I28	职业人士	金融硕士项目负责人	男	31	高（8/12）	模式 A
I29	职业人士	连续创业者、AI 应用公司创始人	男	32	中（6/12）	模式 C
I30	职业人士	AI 驱动电商联合创始人	男	41	高（8/12）	模式 A
I31	职业人士	广告创意专家	女	28	极高（9/12）	模式 A

I32	职业人士	咨询行业创意顾问	女	29	高（8/12）	模式 A
I33	职业人士	量化交易专家	男	31	极高（10/12）	模式 D
I34	职业人士	化学研究员	男	30	高（8/12）	模式 D
I35	职业人士	软件开发工程师	男	24	高（7/12）	模式 A
I36	学生	人文学硕士生	男	33	中（6/12）	模式 A
I37	学生	工商管理硕士生（MBA）	女	24	高（7/12）	模式 C
I38	学生	数据科学与人工智能本科生	男	22	高（8/12）	模式 E
I39	职业人士	工业安全顾问	男	36	高（7/12）	模式 C
I40	学生	航空航天工程博士生	男	26	高（7/12）	模式 A
I41	学生	理论光学博士生	男	25	极高（10/12）	模式 E
I42	学生	材料科学博士生	女	26	高（8/12）	模式 A
I43	职业人士	健身教练 / 内容创作者	男	28	高（7/12）	模式 C
I44	学生	学习科学硕士生	女	25	高（8/12）	模式 E
I45	职业人士	生物医学光学研究员	男	31	极高（9/12）	模式 A
I46	职业人士	AIGC 项目负责人	男	31	中（6/12）	模式 C
I47	职业人士	化工行业销售经理	女	45	中（6/12）	模式 C
I48	学生	电气与电子工程本科生	男	21	中（5/12）	模式 B
I49	学生	化学工程与商业本科生	男	22	中（6/12）	模式 B

表 3-3 呈现三个关键发现。第一，元认知复杂度分布显著分化：10 名受访者达到“极高”水平（9-12 分，占 20%），26 名为“高”水平（7-8 分，53%），13 名处于“中等”水平（5-6 分，27%）。这一分布初步表明，即使在我们有目的性招募的"频繁 AI 使用者"样本中，元认知参与度仍存在实质性差异，而非均质化的高水平。第二，五种有效模式的流行率呈现明显非均匀分布（卡方拟合优度检验：$\chi^2(4)=18.44$，$p=.001$），依次为：模式 A（战略控制型，37%，n=18）为最普遍有效模式，显著超出随机期望（二项精确检验：$p=.005$）；模式 C（情境适应型，33%，n=16）为第二常见；模式 E（元认知教学型，14%，n=7）、模式 D（深度验证型，8%，n=4）、模式 B（迭代优化型，8%，n=4）依次递减。值得注意的是，我们在 49 名受访者中未观察到模式 F（被动依赖型）——这一"缺失模式"本身具有方法论意涵：它反映了自我选择偏差，即愿意接受"AI 使用经验"访谈的用户本身已具备一定元认知觉察。根据 12 位教师/导师的二手报告，被动依赖型用户在学生群体的实际流行率可能达 25-40%，但该群体系统性地回避了本研究（limitation）。第三，也是最关键的发现：人口学变量不预测模式归属。卡方独立性检验显示，教育水平（研究生对比本科）与模式 A 成员身份无显著关联（$\chi^2(1)=1.23$，$p=.27$，Cramér's V=.16），学科背景（STEM 对比非 STEM）同样无预测力（$\chi^2(1)=0.89$，$p=.35$），年龄与模式归属的点双列相关微弱且不显著（$r=.08$，

p=.73）。这些零发现具有深刻理论意涵：它们证伪了专业知识水平可靠预测恰当 AI 使用的假设（Glikson 与 Woolley, 2020），转而支持我们的核心论点——有效使用本质上是元认知成就。

接下来，我们详细阐述六类模式的质性特征。

3.4.2 六类元认知使用模式的深度质性描述

（1） 模式 A ：战略性分解与控制

模式 A 是出现频率最高的有效路径，占参与者的 37%（n＝18）。该类以主动的任务分解、严格的人类监督以及在人与 AI 之间设定原则性边界为主要特征。此类用户并非单纯“使用”工具，而是编排一套经精心管控的人–AI 协作，始终保持 Suchman（2007）所称的“认知主权”，即便这意味着为维护学习的完整性而牺牲效率。

模式 A 在经验材料中呈现三项一致的行为特征。其一是主动的任务分解。用户将复杂问题拆解为可控的子任务，并在各阶段保留监督控制。这不仅是操作层面的考虑，更是一种为在协作过程中设置“认知检查点”的策略性选择。受访者 I1（市场营销博士生）展示了其流程化做法：“我从不把整项任务交给 GPT。我会先把一万字手动压缩到两千字，再逐段投喂给 GPT 打磨。我需要逐段比对 AI 的改动，因为一次给它太多内容容易使其粗暴处理。”值得注意的是，I1 的前置压缩本身即为大量认知工作，她有意不将其外包给 AI，因为该步骤要求把握核心论点，交给 AI 会失去对“何者为要”的控制。此类任务分解在 A 类中出现比例为 89%（16/18），显著高于总体的 45%（22/49；χ^2＝10.67，p＝0.001）。受访者 I3（工商管理博士生）明确表达其原则：“凡是我能做的，绝不交给 GPT。它擅长的是从 0 到 1。”这一“0 到 1”的比喻揭示了 模式 A 的哲学。AI 适合作为空白起点的结构化与信息补充，人类则负责从“1 到 10”的精炼与落地。I3 进一步说明其核验流程：“不熟悉的编码法，GPT 能给基本结构。但我要自己测试、上网交叉核查、找出错误，再把学到的东西带回去提示 GPT。循环是 GPT→测试→在线检索→带着发现再提示。”这种“三角核验”体现出对 AI 输出的审慎立场。

第二项特征是明确的人–AI 角色划分。模式 A 用户维持清晰、并常以生动隐喻加以表述的人机能力区分。受访者 I16（计算机科学博士生）描述了其角色关系的演进，揭示在人保持控制原则的同时分工认识如何随时间走向成熟。他的轨迹呈现四个阶段，体现出角色概念的逐步精细化。第一阶段，AI 定位为从属助手：“我自己写代码的架构，GPT 填函数或完成重复样式。我写整体结构与逻辑，GPT 负责样板部分。GPT 可能贡献了最终代码的 20% 至 30%。”此阶段对应“AI 为助手”的常见框架，人保持主体

性，AI 支持日常性子任务。第二阶段，权责逐步再分配："我开始用详细注释写规范，说明我想要什么，然后让 GPT 生成整段函数。我的角色转向系统架构与代码审查。GPT 产出或许占到 50%，但我会仔细审阅，经常大幅重写。"此阶段引入效率与控制之间的张力，审阅负担在一定程度上抵消了生产率收益。第三阶段，角色明确再定义，I16 以高度清晰的表述总结："我从把 AI 当助手，转向把自己当项目经理、AI 当程序员。我用自然语言写详尽 brief，描述需求、约束与边界情况，然后把大约 80% 的编码委派给 GPT。我的工作是战略规划、质量保证与集成，确保 AI 生成的各部分协调一致并满足规范。"此举标志着根本性再概念化，即由"人亲自编程、AI 协助"转向"人管理、AI 编程"。第四阶段，引入精细化验证基础设施："我现在把同一复杂编码任务同时发给 ChatGPT 与 Google Gemini。如果二者给出的思路相近，我对方案的可靠性有 90% 的把握。如果差异明显，我会追查原因，很多时候是一个模型误解了我的规范，或者确有多条可行路径。并行比对能捕获单一模型不会主动承认的错误。"

关键在于，I16 强调第三阶段的大量委派并未违背模式 A 的控制原则："我委派的是执行而非决策。所有重要的架构选择、算法取舍、性能与质量的权衡，都是我来决定。AI 实现我所规定的内容。若 AI 建议的路径与我不一致，我会否决。我始终处于主导地位。"在理解模式 A 用户如何在不放弃控制的前提下扩大 AI 使用范围时，需把握"执行权可委派，决策权须保留"的区分。同时，I16 为预防过度外包带来的技能退化，主动实施能力维护："我强迫自己每周完全手写解 LeetCode 题，不用 Copilot，不查 ChatGPT，就我和 IDE。我需要在生产项目中转向管理角色的同时维持原始编码能力。这像运动员在无器械条件下做力量训练。"他保留训练日志，记录每周"非辅助编码时长"，设置八小时的最低阈值，将能力维护视为与生产并行的必要训练。类似地，受访者 I20（教育领域研究员）以教学隐喻框定关系："我是老师，AI 是学生。它在语言任务上熟练，但在思考上较弱。我必须全程监督。"受访者 I7（21 岁，本科生，电子商务）则表述为"他擅长检索与写作，我擅长分析。由我决定哪些有用，以及如何修改。"这些分工一致指向人保有决策权而将执行外包给 AI，从而维持我们所称的"战略性监督非对称"。表 3-4 汇总了模式 A 用户的角色划分模型，显示出在隐喻表达差异之下的高度原则一致。

表 3-4　模式 A 用户的角色分工模型（n = 18）

受访者	人类角色	AI 角色	边界维护策略
I1	内容架构师、质量控制者	分段润色者	"绝不进行整任务交接"

I3	能力守护者、验证者	0 到 1 的起始者	“凡是我能做的事，不让 GPT 代劳”
I7	分析者、决策者	信息整合者	“我来决定有用与否及修改方式”
I16	项目经理、质量工程师	程序编写者（约 80%代码）	“详细任务说明与结构化监督”
I20	教师、监督者	学生、执行者	“每一步都要监督”
I28	策略型写作者、事实仲裁者	语言润色者	“先写完整草稿再交给 AI”
I31	创意策略师、品牌守护者	格式限定助手	“先定框架，再让 AI 填充”

第三项特征是多层级验证与精细边界管理。模式 A 用户从不盲目信任 AI 输出，普遍采用系统化核查机制，验证强度平均为 9.0/10（标准差 0.9），显著高于样本均值 5.8/10（$t = 8.93$，$p < .001$）。I3 的三角验证即为典范：“GPT 先生成代码，我自己测试调试，再在线交叉核查，随后带着发现回到 GPT。有时网上方案更好，有时 GPT 的思路更优雅。最终由我裁决。”这种由用户充当最终仲裁者、整合多源信息而非依附单一权威的姿态，是将成熟信息素养运用于 AI 情境的体现。在常规验证之外，部分模式 A 用户发展出高度成熟的边界管理技术。受访者 I31（广告创意专家）实施我们称为“策略性不完美注入”的做法，即在 AI 生成内容中有意引入轻微瑕疵，以维持“真实感”信号。他解释道：“我用 ChatGPT 起草专业邮件，尤其是英文邮件，因为我并非母语者。但我要它故意保留大约 5% 的语法错误或略显生硬的表述。因为我不是母语者，我不希望这封邮件看起来完美无瑕。若同事突然收到我完全无误的英文，他们可能会怀疑使用了 AI 或请人代写，这两种印象对我都不利。”

此做法折射出对 AI 使用所涉社会动力的深刻理解。传统设想认为 AI 的价值在于产出尽可能完美的文本。I31 认识到，当完美与既定社会身份相冲突时，完美反而可能成为负担。他的同事习惯将其视为“能力充足但并非母语者”，突如其来的完美会破坏这种身份一致性，引发怀疑而非赞许。其实施颇为成熟：I31 先令 ChatGPT 生成精炼版本，再明确提示“请将文本修订为包含少量细微语法不当与略显生硬的表述，符合熟练非母语者的写作特征。在保持清晰与专业的前提下将‘完美度’适当降低大约 5%”。获得“非完美”版本后，I31 会复核以确保不完美显得自然而非随机，必要时手工调整，以贴近其常见的个人表达特征。I16 的双模型验证体系在 40% 的模式 A 用户中出现（8/19），将 AI 从单一权威转化为“需经审议的同侪贡献者”。其做法通过并行查询实现，他在 ChatGPT 与 Gemini 中保持同一上下文，同时下发一致规范。当两模型就实现路径达成一致，他会以较高信心推进；当二者分歧显著，他将进一步剖析。有时分歧

揭示了他在规范中的歧义，促使完善规范；有时表明一方误解而另一方理解恰当，从而更新后续的信任校准；偶尔二者给出各有取舍的有效路径，则需要他明确更重要的约束条件。

而后，将模式 A 的行为映射至 Flavell（1979）的元认知框架，可见四类过程均有稳定证据。规划过程在 95% 的个案中表现突出（17/18），体现为明确的任务分解、策略性角色定义与对 AI 介入时机的目标导向选择。监控过程在 89% 的个案中明显（16/18），体现为持续的过程追踪（如 I1 的“逐段比对”）、实时质量评估（如 I16 的双模型体系）与主动的错误探测。评估过程在 89% 的个案中显著（16/18），体现为对输出质量的批判性评估、对能力边界的认知（I3 的“0 到 1”模型）与据以进行的信任校准。调节过程在 72% 的个案中呈中到强的证据（13/18），体现为基于监控结果的动态策略调整，以及在验证揭示不足时进行工具切换。I16 的四阶段演进显示，模式 A 用户在“元层级”也持续开展调节，不仅调整即时战术，而且重构整体策略路径。其第四阶段的双模型验证可被视为对“验证过程本身”的调节，即在保持人类裁决权的同时，逐步将验证由纯手工转向体系化的多模型验证。

量化分析进一步凸显模式 A 的独特性。其元认知复杂度均分为 8.1/12（SD = 1.1），显著高于样本均值 7.5（$t = 3.84$，$p < .001$）。验证强度 9.0/10 显著高于一般实践。任务分解发生率为 89%，高于总体的 45%。尤为重要的是，人口学变量与模式 A 归属不显著相关：教育层次（博士与本科）$\chi^2 = 1.23$，$p = .27$；学科背景（理工与非理工）$\chi^2 = 0.89$，$p = .35$；年龄相关 $r = .08$，$p = .73$。此类“零结果”具有重要理论含义，表明模式 A 的高水平元认知做法超越传统用户分类。受访者 I7（本科生，电子商务）尤能说明此点。尽管资历尚浅且受训有限，I7 的模式 A 成熟度与年长博士生相当甚至更高。他的表述“他擅长检索与写作，我擅长分析。由我决定哪些有用，以及如何修改”，呈现出与资深用户一致的边界设定与判断保留。相反，我们亦观察到部分博士生在元认知策略上明显不足，甚至趋向模式 F 的过度依赖。教育层次内部的此种差异进一步强化了本研究核心论点：决定成效的是使用模式，而非用户身份。

模式 A 的行为揭示了四项关键系统设计需求。 首先，针对当前系统常自我定位为“可处理整任务”从而无意鼓励模式 F 的被动接受，系统需要提供任务分解脚手架。当用户输入的任务超过复杂度阈值时，系统应主动提出分解方案，帮助用户将复杂任务拆解为可管理的阶段。I1 在每次交互前花费 40 分钟手动压缩文档的经历表明，尽管 89%的模式 A 用户主动分解任务，但这一过程耗费大量认知资源，系统化支持可显著

降低门槛。 其次，I1 关于“高亮编辑内容”的诉求揭示了过程透明性与可追溯性的缺口。模式 A 用户需要明确呈现“改了什么、改在何处、为何修改”。系统应提供修订模式、并排前后版本对照与带差异可视化的版本历史，通过将 AI 的变换过程外显化来支撑用户进行逐段级监控。 第三，需要将“AI 为工具而非替代”的理念落实为人类能动性保护的设计默认。系统建议应以选项而非指令呈现，提供多个方案而非单一“最优答案”，并明确标识“此为 AI 的建议，是否采纳由您决定”。关键在于，系统应对独立完成的工作给予正向反馈，而非仅奖励 AI 协助带来的效率。I31 的“策略性不完美”实践提示，系统还应支持“有界优化”，允许用户指定与其真实能力相匹配的目标质量水平。 第四，I16 的“我是经理、GPT 是程序员”、I20 的“我是教师、AI 是学生”表明，有效协作建立在清晰的角色定位之上。然而，19 名模式 A 用户中有 39%反映缺乏显性机制来协商角色边界。系统需要提供角色定义指导，允许用户选择协作模式（如教学模式、审查模式、咨询模式等），并在界面显著标识当前角色配置。I16 的四阶段演进表明角色关系非静态，系统应支持用户根据能力发展动态调整分工。

模式 A 对人机协作理解作出三方面贡献。其一，将交互记忆系统（TMS，Wegner，1987）扩展至人机二元关系，并揭示关键差异。TMS 关注群体如何通过“谁知道什么”的分工以效率最优地委派任务。然而，模式 A 用户在效率受损时仍坚持原则边界。I3 的“凡我能做者不交 GPT”明确背离 TMS 的效率逻辑，提示一种新概念：人机协作中的“认知主权维持”。人们维护能力不仅出于效率考量，更关涉身份守护与技能维持，动机超出工具理性的范畴，涉及发展性与专业身份。其二，对近年来“AI 为合作伙伴”的表述提出复杂化（Seeber 等，2020）。此类用户明确拒斥平等伙伴定位，而建立层级关系：I20 自身定位为教师、AI 为学生；I16 为经理、AI 为程序员；I3 将 GPT 定位为“从 0 到 1 的工具”，而从 1 到 10 的执行由人掌控。此种“人监督、AI 从属”的层级定位，在学习语境中可能较“伙伴平等”更为健康，后者容易诱发对 AI 建议的过度依赖。I16 的四阶段演进显示，层级关系并非静止，而是随着用户经验与 AI 能力的变化被动态协商。由“AI 协助人类编程”过渡到“人类管理 AI 编程”，表面权力结构发生变化，但其深层结构未变，即无论执行如何委派，人始终保有决策权。这提示有必要在理论上区分两类协作：以生产效率为导向的协作与以学习能力发展为导向的协作。前者允许更为平行的权责分配，后者则要求明确的人类主位，以防止能力退化并维持发展性投入。其三，从用户侧赋予“负责任使用 AI”的实证内涵。模式 A 从用户侧为“负责任使用 AI”提供了经验证据。尽管伦理讨论多集中于开发者侧的公平、透明与问

责，模式 A 用户展现了用户侧的责任实践：先验证后信任（认识论责任）、能力保存优先于便利（发展性责任）、基于原则而非机会主义的外包（诚信责任）。I31 的“策略性不完美”为此增添了社会性责任维度，即认识到 AI 使用发生于社会语境，需关注他者感知与交往后果。其对“同事如何解读突然完美的英文”的敏感性表明，负责任使用不仅是个人任务达成，更涉及对社会关系的审慎管理。这为既有较为抽象的“实践中的 AI 素养”（Ng 等，2021）提供了具体行为层面的支撑，表明成熟使用根本上体现在具体的元认知策略，而非仅是知识性掌握。I16 由第一阶段迈向第四阶段的过程也说明，负责任实践源于迭代经验、策略性试验与持续校准，而非一次性的指导或规则遵循。

（2）模式 B：迭代优化与校准

模式 B 的主要占比虽低，仅 8%（n = 4），另有少数受访者在特定情境下呈现模式 B 特征，却揭示了用户如何通过“容错性坚持”来应对 AI 的不可靠性。与模式 A 的“控制优先”路径不同，模式 B 接受甚至拥抱失败作为发现过程的一部分，与学习科学中的“生产性失败”（Kapur，2008）理论相契合。

区分模式 B 的核心特征在于：即便明知 AI 在当前任务上的局限，仍持续与之交互。受访者 I9（管理科学本科生）概括道：“我知道它做不到……我知道 GPT 的输出不会完全符合我的需求，但我会不断尝试。它不会承认做不到，而是会给出看起来对但并不对的东西。”被问及此种“非理性坚持”的动因，I9 解释：“每次失败都会让我学到点什么。到第十五轮时，我终于拿到了 70% 的内容。剩下 30% 我自己补，但此时我已更理解问题。”此种将失败视作迭代精化契机的重构，与模式 A 中两三次失败即转为纯人工路径的做法形成鲜明对比。这种坚持常表现为多轮校准。I16（在其他情境亦呈现模式 A 特征）提到其最长项目“超过 50 轮才把代码调好。每一轮都基于出错点来微调提示。”该过程通常呈三阶段：第一阶段为初始提示，往往出于探索而设置较宽，得到错误但具信息价值的输出；第二阶段为错误分析与提示精炼，依任务复杂度而重复 5–20 次；第三阶段达成“足够好”的解，虽不完美，但在时间投入下实现了足额价值获取。I15（23 岁，战略分析）描述了“双重优化循环”：“我先手工精炼草稿，确定薄弱处，再交给 GPT 并要求‘基于清晰、简洁、专业性优化’。然后把我的版本与 GPT 的版本对比，取其所长。就像有两个编辑。”模式 B 的另一特征是策略性“模型切换”。当某一模型失利，用户并不放弃“基于 AI 的路径”，而是切换模型。I5（23 岁，数据分析硕士生）表示：“若 ChatGPT 报错或明显错，我会把同一提示复制到 DeepSeek。有时 DeepSeek 在中文或近期事件上更有效。”与模式 A 的“针对模型的信任校准”以及模式

D 的“并行多模型验证”（同时比对以作验证）不同，模式 B 更像把不同模型视为同类工具箱中的替换件，按序列备用。

相较模式 A 的“前置规划”，模式 B 呈现独特元认知画像：规划过程证据偏弱，仅 40% 个案明确体现，常见为探索式目标设定（“先看看会怎样”）；监控过程则非常强，达 95%，表现为跨迭代的持续质量核查与细致的限度识别（“我知道它错，但错到什么程度”）；评估过程较强，为 70%，表现在从瑕疵输出中提取部分价值的能力，以及对连续版本的比较判断；调节过程非常强，达 90%，表现为高度适应性，能在中途彻底重构路径，并平静地执行工具切换而不伴随挫败情绪。与模式 A 的关键差异在于时间分布：模式 A 将元认知前置于规划环节（“预防错误”），模式 B 则将其后置于监控与调节（“接受错误并从中学习”）。尽管作为主要模式的占比最低（8%，n=4），模式 B 用户通过持续迭代能够完成复杂任务，但完成同等任务所需时长通常显著长于其他模式。其元认知复杂度均分为 6.0/12，低于模式 A 的 8.1，但分布结构不同：规划弱而调节极强，说明"有效性"可在不同元认知维度上呈现。模式 B 在三类情境中尤为有效。其一，目标尚不明确的“弱定义问题”，迭代有助于经由渐进式精炼来澄清需求。I9 的素材生成即属此类：“我本来就说不清要什么，直到看了 GPT 的多次尝试。每个版本都让我更明确要求。”其二，创意任务中，用户有意利用他人视为“AI 失败”的产物。I13（AI 工程方向）指出“幻觉”可作为“创意资产”：“头脑风暴时，我希望 GPT ‘幻觉’。它会给 20 个疯狂点子，其中 18 个是无稽之谈，但有 2 个非常好，是我自己想不到的。这就够了。”其三，快速原型情境中，I16 的 50 轮编码迭代比从零手写更快，即便单轮效率并不高。

模式 B 启发三项以迭代为中心的设计需求。 首先，当前系统使迭代代价高昂，表现为上下文丢失、需手工复制粘贴、难以跨版本比较。系统需要提供低成本迭代机制，包括自动版本控制并保存每次迭代及时间戳、相邻版本差异高亮、一键回滚任意旧版、任意两版并排 A/B 比对、以及展示提示演化与输出变化的“迭代图”。 其次，回应用户在 ChatGPT、Claude、Gemini、DeepSeek 之间手工迁移提示的现实，系统需要支持跨模型实验。具体包括单一提示可同时发往多模型的界面、基于用户投票的输出排序、依据任务特征的模型推荐，以及按任务类型记录“哪类任务用哪模型更佳”的个性化学习系统。 第三，需要建立失败容忍与学习机制，将错误由“问题”转化为“学习机会”。包括失败分析以解释失败原因、失败样本库以保存失败尝试供将来参考、鼓励性信息以将迭代定位为成熟使用、以及“边界觉察”在多次失败后提示“该任务可能超出当

前 AI 能力”。I9 坦言曾耗时数小时强迫 GPT 完成本可 20 分钟手工完成的任务，提示系统需在鼓励"生产性坚持"与防止"无效执拗"之间取得平衡。

模式 B 将“生产性失败”理论由人类学习拓展至人机协作。其一，在创意情境下对“AI 幻觉”进行再诠释，将其视为潜在的生成资源（如 I13 的头脑风暴）。其二，挑战以效率为唯一评价标准的做法——模式 B 更“慢”，却拥有更高的最终成功率，表明“有效性”与“效率”是需分别优化的两个维度。其三，指出“迭代本身”是一项元认知技能，高迭代用户的“调节型元认知”显著更强，提示迭代可培育可迁移的适应能力。然而，模式 B 亦存在需系统层面化解的风险。“坚持谬误”即在转为纯人工更快更有效时仍继续与 AI 纠缠，为其主要局限。I9 的自述即为明证。设计上需通过 MR7 所述的边界觉察帮助用户识别“应当止损”的时点。

（3）模式 C：情境敏感的适配

模式 C 由 33% 的受访者呈现（n = 16），体现了一种根本不同的人机协作取向，即基于任务特征而非稳定的个体倾向进行动态信任校准。此类用户会根据具体情境采用截然不同的策略，包括高风险与低风险任务、熟悉与不熟悉领域、创造性工作与事实性工作、可核查判断与主观性判断的差异。本研究对人机信任领域的既有假设构成关键挑战（Glikson & Woolley，2020；Lee & See，2004）：同一位个体可能在某一情境几近于零信任，而在另一情境完全信任，且在同一用户不同任务类型之间的差异有时可超过 100 个百分点。

首先是基于风险的信任分层与情境校准。最为鲜明的案例来自受访者 I2（会计学博士生），其呈现“反向信任非对称”。I2 明确表示：“在学术写作中，我坚决不使用 ChatGPT 的任何部分：不做句子润色，不做构思辅助，不做翻译。原因包括期刊明令限制、存在检出风险、语言风格模板化、原创性要求高以及数据隐私担忧。”这在高风险学术场景体现为“零信任”。然而在个人情境下，I2 又表示：“我完全信任 GPT 来算命。我输入我的出生时间八字，让 GPT 扮演中国算命先生。我确实会相信它说的。”当被追问这种“矛盾”时，I2 给出了深层的理由。该受访者从多维风险评估进行解释：“高风险任务需要完全可控。低风险任务得益于它的视角——即便错了也无妨，还有可能有启发。这是一种简易的‘抵抗不确定性’方式。”由此，AI 不再是“统一可信或不可信”的对象，而是需根据后果展开衡量的资源。表 3-5 将 I2 在五类情境中的信任谱系进行映射，显示在同一时期内信任水平可从学术写作的 0% 到算命占卜的 100% 全幅度变化。

表 3-5 受访者 I2 的情境依赖型信任校准

情境	信任水平	核验	利害相关	理由
学术写作	0%	不使用	极高	“存在检出风险；要求原创”
文献检索	0%	不使用	高	“存在引文幻觉”
早期研究大纲	85%	最小限度	中高	“我们一开始都无从下手”
英语语法检查	65%	反复口头确认	中	“需保证专业交流”
占卜与解签	100%	无	低	“抵抗不确定性；即便错误也无害”

此模式在其他受访者处亦有体现，具体情境与分布有所差异。I4（会计学硕士生）陈述了明确的决策算法：“超过 100 页的论文我手工做。30 页的课堂作业我用 AI。我的计算是：任务重要性 × 预期 GPT 质量 × 修订成本。”这表明其情境敏感不是“凭直觉”，而是系统化的。I4 指出，随着实践积累，这一计算“已变得自动化，我能直观感知何时值得用 AI”。I10（37 岁，金融副教授）以更直白方式表述这种“反向关系”：“ChatGPT 处理垃圾活。任务越重要，我越不信它。”模式 C 用户通常维持一个“策略储备”，即根据任务快速分类来调用不同的既定方案。I4 的成本收益公式表现为多节点的程序化决策：若任务重要性为“高”，默认为手工；若长度超过 100 页且重要性为“中”，仍然手工；若基于类似任务经验判断 AI 预期质量为“低”，则手工；其他情况下计算投资回报率为“节省时间 ÷（修订时间 + 风险成本）”，若比值超过 2.0 则动用 AI，否则手工。这一多阶段决策树最初需要有意识计算，但经反复应用已逐步“程序化”（Anderson，1982）。

“领域敏感的信任分层”也是模式 C 的特征之一。I43（健身教练）概括了“专业悖论”：“在我的领域（健身），AI 就像初级同事。我能立刻抓出错误。但在我不熟悉的领域（如平面设计），AI 几乎可以替代。只有需要生活经验的地方才离不开人。”在这一关系中，用户的领域专长越高，对 AI 的信任反而越低；专长越低，信任越高。这与“专家因校准更好而更信任 AI”的线性假设相悖。事实上，专长带来的是错误识别能力，而一旦识别出错误，信任自然下降；相反，初学者在陌生领域适度信任 AI 以弥补知识缺口则可能恰当。I43 进一步阐释健身领域的经验：“当客户向 GPT 请求训练计划时，我能马上发现问题：忽略关节健康的进阶安排、会导致过度训练的训练量建议、可能导致受伤的动作提示。GPT 的回答像教科书，但忽视个体情境，例如既往伤病、动作受限、训练史、心理因素。GPT 的建议中或许只有 40% 可直接使用，其余 60% 需

要专家修订。”这种低信任源于 I43 能凭专长识别 AI 的局限。该受访者将之与平面设计经验对比：“制作营销材料时，我让 GPT 生成设计。我无法像专业设计师那样识别细微缺陷。或许 70% 的结果对我的目的‘足够好’，剩下 30% 的不足我也意识不到。就功能而言，GPT 的 70% 已超过我自己的 30% 能力。”此“专长悖论”产生反直觉预测：在某一领域，随着用户专长提升，最优的 AI 依赖度可能下降而非上升。新手适度依靠 AI 弥补差距，中等水平用户进行协作性增强，高水平专家则因可识别错误而降低依赖。这一“U 型关系”挑战了关于专长与信任线性关系的假设。

受访者 I18（航天研究员）针对六种不同 AI 模型维持显式信任谱系，按用途分别部署：“我用 ChatGPT 解答一般问题，用 Claude 进行写作，用 Gemini 查阅学术文献，用 Perplexity 获取时事并带来源引用，用 DeepSeek 处理中文语境，用 Cursor 融入编码流程。通用任务上的信任超过 80%，在学术工作上低于 60%。差异化信任至关重要。”此多模型策略反映出对模型间训练数据、优化目标与由此产生的强项差异的成熟认识，类似于人类协作研究中的“交互记忆”意识（Wegner，1987），但在此扩展至人—AI 工具生态系统。I18 的差异化信任来自系统性试验：针对不同任务类型测试各模型，记录最优适配。ChatGPT 在会话与常识方面表现优良，但偶有信息时效不足；Claude 在写作质量与风格一致性上更佳；Gemini 在新近学术文献方面更为便利，但有时误解技术性提问；Perplexity 专长于时事并注重来源标注；DeepSeek 在中文语境下的自然度更高；Cursor 与编码工作流整合最佳，尽管偶尔建议过时方法。I18 并不将任何一个模型视为“最好”，而是维护一套以任务特征映射模型选择的专门化信任画像。

受访者 I26（电商企业家）将 AI 定位为“虚拟联合创始人”，同时保持严格的领域边界：“我的工作流程总以 AI 为起点，它是顾问、导师，也是个体创业时的反馈替身。但关于我特定客户群或竞争优势的问题我从不问它。市场才是最终验证者。”这体现了对专有知识边界的认识，即 AI 无法拥有的企业特定信息，同时也承认 AI 在训练数据可及的通用商业策略层面的价值。I26 的边界设定并非简单的高低信任，而是在通用商业框架上保持高信任（80–90%），在公司特定战略决策上保持零信任或近零信任（0–10%），明确指出 AI 缺乏关于其市场地位、客户关系与竞争动态的专有信息。

模式 C 在元认知上呈现显著强势的“评估过程”。规划过程为中等水平（63%），其核心在于基于快速任务分类进行情境化策略选择，而非模式 A 式的前置详尽规划。监控过程亦为中等（62%），更多聚焦于基于领域的信任校准与持续风险评估，而非输出层面的细粒度核查。评估过程非常强（90%），表现为对任务风险、复杂性、错

误成本与时间压力的多维分析、显性或隐性的 ROI 计算、以及对“何时助益何时妨碍”的能力边界识别。调节过程较强（71%），体现为随情境变化进行灵活的工具与策略切换，且多发生在任务之间而非任务之内。区分模式 C 的关键元认知技能是快速任务分类，即能沿风险水平、领域熟悉度、可验证性、时间敏感性等多维度即时对新任务进行归类，并从策略库中检索相应策略。这与 Ericsson 与 Lehmann（1996）所谓“编译化专长”一致，依赖模式识别实现无需意识层面的快速准确决策。I4 便是例证，该受访者最初需要显式计算的成本—收益公式，经反复应用后转化为自动化的直觉，使其无需刻意分析便能“感知”何时值得使用 AI。这一程序化过程带来认知效率提升，源自模式 C 的实践，体现为由耗力的深思发展为流畅的直觉。

量化结果表明，模式 C 是第二常见的有效模式（33%，16/49）。同一受访者在不同情境的信任标准差平均为 34.2%（范围 18%–87%），显著高于模式 A 的 8.3%（$F = 89.3$，$p < .001$），证实“情境敏感性”为其定义性特征。模式 C 用户平均涉入 3.8 个不同情境（标准差 1.6），其策略储备平均包含 4.3 种不同做法（标准差 1.8），显著高于模式 A 的 1.9（$t = 4.23$，$p < .001$）。尽管其总体元认知复杂度为中等（6.4/12），但在“评估”分量表上得分最高（8.7/10），体现“判断与校准”对该模式的重要性。关于模式稳定性，我们得到一项关键发现：模式 C 用户并非不一致，而是精准的情境校准。访谈数据显示，模式 C 用户在相同情境下倾向于采用一致的策略，而不同情境之间的信任随意图而系统性变化。这表明看似“不一致”的人机信任实际上是成熟的情境依赖性校准，即用户在情境内保持稳定策略，同时在情境之间灵活调整，反驳了将信任视为稳定个体特质的研究（Glikson & Woolley，2020）。该稳定性发现具有理论关键性：它区分了模式 C 的策略性适应与模式 F 的不可靠摇摆。模式 F 用户即使在同一情境内也重测可靠性低，其信任与使用更大程度受瞬时便利、近期经历或情绪驱动，而非系统的任务分析。模式 C 用户则在情境内具备高可靠性，并在情境间呈现有意的差异。此“稳定中的灵活”是适应性专长的标志（Hatano & Inagaki，1986），即在熟悉情境中保持一致策略，在新情境中灵活调整。

模式 C 启发三项面向情境敏感性的设计需求。 首先,针对现有系统一视同仁地对待所有查询的现状,系统需要提供任务特征识别能力。应对输入任务进行自动分类,涵盖关键维度：风险水平、可验证性、领域维度、错误成本等。系统输出需提供透明分类与建议，例如“已将此任务分类为： 高风险（学术投稿）、专长外（生物化学）。建议：提高验证强度”。I2 的五情境信任谱为该需求提供经验证据，其学术写作与算命场

景的信任差异达 100 个百分点。 其次，针对“单一信任等级”假设的不足,系统需要支持动态信任校准。系统不应仅询问“你是否信任该 AI”，而应展示情境特定的信任信息：检测到的任务类型、用户在类似任务中的历史信任、相似用户的信任水平、AI 在本次查询中的自我置信，以及信任不匹配的识别。实现上应跨情境×领域×任务类型维度跟踪信任，并呈现与当下任务相关的子集。 第三，需要提供成本效益决策支持，将 I4 的 ROI 公式嵌入为决策辅助。邀请用户输入任务重要性、纯手工完成时间、AI 辅助与修订时间、对该任务类型的 AI 置信，随后给出手工与 AI 辅助的期望时间对比，并纳入“需完全重做”的重大错误的概率化考量。实现上可加入结果回溯学习，构建个性化预测模型，支持较少经验的用户形成与模式 C 类似的校准能力。

在理论层面，C 类贡献体现在三方面。其一，挑战了将信任视为稳定个体特质的主流模型，模式 C 从根本上挑战了人机信任研究中将信任视为缓慢随经验积累而演化的个体特质的假设（Glikson & Woolley，2020；Mahmud 等，2022）。模式 C 显示信任应被视为状态，而非特质，即具有情境性、多维性与快速可切换性。这要求修正理论模型。传统方法将 Trust（Person，AI）视为单一随时间单调演变的值。模式 C 揭示信任为动态张量：Trust（Person，AI，Task，Context，Time），各维度相互之间呈乘法式而非加法式作用。其二，跨情节的适应性元认知与策略选择。模式 C 将适应性元认知从其常见的“情节内”应用扩展至“情节间”。Azevedo 与 Hadwin（2005）的自我调节学习框架主要描述学习情节内部的适应，即监控理解并在任务中调整策略。模式 C 进一步展示用户在任务之间发展条件性生产规则：若（任务＝学术写作）且（风险＝高），则（策略＝零 AI，验证＝不适用）；若（任务＝构思）且（风险＝低），则（策略＝充分使用 AI，验证＝最小化）。这些规则代表编译化的元认知专长，即 Ericsson 与 Lehmann（1996）所称的“块”，在此应用于 AI 使用策略选择。其三，通过维度化刻画化解表面上的信任悖论。模式 C 解释了文献中的“信任悖论”现象，即同一用户同时报告“高信任”与“低信任”（Lee & See，2004）。模式 C 说明这并非悖论，而是精确性问题：信任不仅依赖个体，也依赖问题。I2 同时对学术写作保持 0% 信任、对占卜保持 100% 信任，是在不同目标（认识论严谨性与情绪调节）与错误成本（职业损害与零后果）下的理性校准。只要在完整的维度空间中对信任进行恰当刻画，表面矛盾便不复存在。

（4）模式 D：深度核验与批判性介入

作为主要模式出现在 8% 的受访者中（n＝4），另有 14% 的受访者呈现次级的 D 特征（n＝7）。模式 D 代表了人机协作中在认识论层面最为严格的一种路径。其核心特征是将系统性怀疑、跨多来源的并行验证以及多方法三角互证设为默认实践，而非偶发性的核查。模式 D 的使用者并不将人工智能的输出视为可直接接受的答案，而是视为有待经由经验性检验的假设。Rader 等（2018）将此称为“参与式核验”，即由用户主动建构可信度，而非被动接受系统给出的置信分数。

第一个特征为并行解题与独立核验，模式 D 的范式性行为是我们所称的“并行解题”，即人在获得 AI 协助的同时独立开展求解，以便进行不受污染的对比。受访者 8（I8，本科，计算机科学）描述其默认流程：“我会与 GPT 并行求解。先独立完成我的方案，再看 GPT 的输出。如果两者一致，当然很好；如果不同，我就追问：谁错了，为什么错。这有两点益处：第一，可以立刻捕捉 GPT 的错误；第二，有时 GPT 的思路更优雅，我也能从中学到新东西。”这种并行方法需要额外投入。I8 报告其用时约为直接接受 GPT 输出但不做核验的两倍，但由此可避免较为宽松做法中常见的虚假自信。更为关键的是，I8 指出了其他模式容易忽视的一种人机协作风险，即过早的认知锚定。该受访者解释说：“看过 GPT 的解法之后，我原本的灵感碎片会消失。如果 GPT 采取某一路径，就会锚定我的思路。因此我会先独立求解，甚至在动手时直接把 GPT 关掉。”这一观察与认知心理学关于锚定效应的研究一致（Tversky 与 Kahneman，1974），表明在独立尝试之前先接触 AI 方案，会以既定的解题轨迹约束人的创造性。模式 D 通过在完成人的方案前保持独立性来避免这种锚定。

多模型交叉验证也是模式 D 的一个标志性行为，但其实现方式不同于模式 B 的序贯式模型切换。受访者 I22（深度学习博士生）发展出其所谓“赛马机制”：“我把同一个提示同时发给 ChatGPT、Gemini 和 Claude。如果三者一致，我的置信度大约达到 85%，虽然仍非 100%，但已很高。如果两个一致、一个不同，我就检查异常者。如果三个答案都不同，那么问题要么超出了当前 AI 的能力边界，要么是我的提示存在歧义。”这一策略在总体样本中出现于 14% 的个体（7/49），其实质是将 AI 从单一权威转化为“证据融贯”的指示器，即自信来自不同系统之间的独立一致，而非某一系统输出的质量或其自我表达的确定性。该方法背后的理论原则可追溯至科学哲学中的“归纳的一致性”（Whewell，1840）。由多条独立证据线索支持的结论，较之单一来源支持的结论更值得信赖，且与来源的个体质量无关。当多个人工智能系统在不同训练数据、

架构与优化目标之下仍然给出相同答案，这种收敛比某一“名义上更强”的单一系统的响应更具说服力。I22 将“三模一致”量化为 85% 的置信阈值，而对单一模型响应的典型置信度仅为 50% 至 60%，由此直观呈现了该认识论原则。受访者 19（项目主管）将三角互证扩展至人的判断：“对于工作中复杂的人际问题，我会询问三方来源：（1）ChatGPT 的分析，（2）导师的建议，（3）同事的看法。我并不把 ChatGPT 当作决策者，而是作为审议过程中的一个声音。”这种人机三角互证将 AI 定位为平行的意见贡献者，而非最高权威，与组织理论中的分布式决策理念一致（Hutchins，1995），即在多主体之间汇聚专业性，而非集中于单一决策者。

第三个特征为主动约束与信息沙盒。除了事后验证，一些模式 D 用户实施主动控制策略，约束 AI 的信息空间以确保可靠性。受访者 I34（化学博士生）开发了“信息沙盒策略”，即先发制人地将 AI 知识库限制在可信来源，再允许其处理研究问题。I34 解释道：“我不会直接把研究问题扔给 GPT 期待准确的文献综述。相反，我遵循结构化协议。”I34 的四阶段协议旨在建立有界信任。第一阶段：人工文献策展。手动从顶级化学期刊识别 4-5 篇高度相关的同行评审论文，通过 Web of Science 等数据库进行 2-3 小时检索，应用领域专业知识识别开创性和高质量研究，下载全文 PDF。第二阶段：实施沙盒。“我上传 PDF 或粘贴摘录，说：‘我提供几篇关于[主题]的论文。请阅读并总结这些文章中的关键发现、方法论和理论框架。’”这将 AI 响应空间限制在提供文档而非一般训练数据。I34 强调：“我需要给它可阅读的文档，让它首先只基于该上下文。这样我确切知道其信息基础，并可在信任其执行更广泛任务前验证其理解。”第三阶段：测试理解。“总结后，我询问关于方法论细节、理论论证或实验结果的具体问题。如果准确回答（我可验证，因我也读过），我就知道它真正理解内容，而非生成听起来合理但可能不准确的响应。”这具有双重目的：验证 AI 阅读理解能力，识别幻觉倾向。第四阶段：谨慎扩展。成功测试后：“一旦确认它正确理解基础论文，我会问：‘基于这些框架，请搜索过去 3 年关于[特定方面]的其他相关论文。’”即使扩展阶段也保持验证：“我对照实际数据库检查每个引用。如果甚至一个引用是捏造的，我会完全停止使用 AI，返回手动方法。”

I34 的信息沙盒的理论意义在于其对典型人工智能使用模式的倒置。标准做法将人工智能视为拥有来自训练数据的全面知识，向其提问并希望响应来自该知识的准确部分。I34 认识到验证人工智能整个知识库的不可能性，从而创建了一个有界的、可验证的子集（精选论文），并约束人工智能在该空间内运作。这代表了我们所称的“有界

信任”，即不是试图校准对人工智能全部能力的信任（不可知）,而是创建信任变得可验证因而合理的有限情境。该受访者将此与同事对比：“实验室其他博士生无约束使用 ChatGPT 做文献综述。他们得到自信的摘要，引用看似相关的论文。但试图访问时，许多论文不存在，出现捏造作者、虚假 DOI 等现象。有些人在 AI 生成的虚构基础上构建研究框架，浪费数周。看到他们困境后，我开发了沙盒协议。它前期更耗时，但防止灾难性失败。”时间投资证明是可观的，I34 估计其的四阶段协议需要 5-6 小时才能完成全面的文献综述，而如果盲目相信人工智能则需要 2-3 小时。然而，与其同事追寻虚构参考文献而损失的数周时间相比，该受访者主动谨慎被证明是高度成本效益的。

第四个特征为专业情境中的系统性怀疑与不可协商的可解释性。在高风险专业领域，模式 D 用户提出了超出当前可解释人工智能研究常规手段的刚性可解释性要求，后者通常局限于置信分数或特征重要性可视化。受访者 I33（量化交易从业者）给出了本研究中最为坚定的立场：“可解释性是不可协商的。在交易中，黑箱意味着不可控风险。我需要知道这笔收益的来源，是散户恐慌，还是机构再平衡。GPT 不能告诉我这一点，所以我不会用它来生成策略。我只在日常代码上用它，而且每一行我都必须搞懂。”对 I33 而言，拒绝黑箱属于职业伦理而非偏好问题：“在交易中使用不可解释的 AI 违背信义义务。如果我无法向客户解释其资金为何发生某种变动，就未尽职责。”这种伦理化的表述指出了在零和竞争领域采用 AI 的根本障碍。被问及强绩效是否能抵消黑箱顾虑时，I33 明确回答：“不会。即便一个不可解释的算法连续五年年化回报 20%，我仍不会用它。一旦它失效，而所有策略终将失效，我将无从知其因、无从修复。这不符合风险管理的基本要求。”受访者 I17（金融学博士生）提出其“误差成本决策模型”：“从概率模型角度看，GPT 很难达到 100% 正确。黑箱特性使得完全信任不可能。我只在错误代价低的场景使用 GPT，对于高风险的金融建模我会手工完成。信任程度与任务风险成反比。”这一定规则强调，随着任务风险上升，可接受的 AI 不透明度下降，直至出现阈值，即便高绩效也因不可解释性而不可用。这一视角不同于传统聚焦于性能与信任关系的校准研究，模式 D 用户将可解释性作为可独立否决性能考量的维度。

主动错误测试是模式 D 的第五类行为，对应 Bansal 等（2021）提出的“校准性体验”，即在不确定问题之前，先用已知答案的问题进行刻意测试以评估 AI 的可靠性。受访者 I42（材料科学博士生）描述其系统化流程：“我用自己有把握的问题测试 GPT，例如基础热力学或我亲自测量过的材料性质。如果它在这些问题上表现良好，我才在

相邻未知问题上谨慎使用。如果它在已知问题上出错，我在该主题上就完全不用。”该受访者将此流程形式化为个人测试规程：第一步，在本领域用 10 个已知答案问题进行测试；第二步，计算准确率；第三步，若低于 80%，则完全拒用；第四步，若 80% 至 95%，仅用于构思并辅以核验；第五步，若高于 95%，则在抽查下谨慎使用；第六步，每月复测，以应对系统更新导致的概念漂移。该规程体现了在一般用户中罕见的经验主义认识论实践，并认识到 AI 的可靠性具有领域特异性（一般题表现无法预测专门题表现）、时间敏感性（系统更新会改变可靠性）与用例依赖性（构思与定稿对可靠性阈值要求不同）。按月复测则反映了对系统演化的敏感，由此要求持续校准，而非一次性评估。

模式 D 的元认知特征在本数据集中呈现出最高的监控强度。98% 的用户（4 位主要 D 用户全体，加之 7 位次级 D 用户中的 6 位）具有卓越的监控行为。规划过程表现为中等强度（65%），其重点在于“核验基础设施”的预先筹划，包括确定交叉核对的来源、拟采用的验证方法与渠道等，而非模式 A 所强调的任务分解式规划。监控过程的投入达到异常水平（98%），体现在连续的质量检查、系统性的交叉验证与高度的错误敏感性。评价过程表现很强（88%），体现在能力判断、刚性可解释性要求以及对“置信与真实可靠性一致性”的成熟评估。调节过程表现为中等（53%），相较其他有效模式略低，其原因在于一旦核验体系建立，后续对策略的频繁调整需求下降，模式 D 更像是一种稳定可复用的系统化路径，而非高度情境适配的路径。模式 D 与模式 B 的关键元认知差异在于时间配置策略。模式 D 将监控“前置”，通过搭设基础设施来实现（并行求解、多模型接入、已知答案测试）；模式 B 则将监控“后置”于迭代过程中。两者均实现核验目标，但路径分别为主动的系统性怀疑与反应性的迭代改进。

统计分析显示，模式 D 是最不常见的有效模式，主要占比为 8%（4/49），次级特征占比为 14%（7/49）。然而低频并不代表低重要性。模式 D 用户高度集中于错误代价极高的领域。所有模式 D 用户的核验强度均达 10/10。以“表达的信任”与“实际可靠性”之间的平均绝对误差衡量的信任校准准确度为 3.2%（为全样本最佳），而样本均值为 18.4%（$U=12$，$p<.001$）。在职业情境中，50% 的主要 D 用户（2/4）来自零和竞争领域，包括金融与交易。AI 错误的检出率达到 94%，而样本总体为 52%（$\chi^2=18.3$，$p<.001$），表明模式 D 的高强度核验可显著提升错误捕捉能力。总体占比虽仅 8%，但在竞争性领域的高集中度（75%）提示模式 D 主要在环境风险足以“逼出”严格实践时才会形成。I33 提到，其公司曾在一名初级分析师使用不可解释的 AI 生成的策

略于市场波动中失效并造成 5 万美元损失后，全面禁止在交易策略中使用 GPT，这一案例直观说明在高风险领域缺乏模式 D 式核验的后果。

模式 D 为三个关键设计要求提供依据。 首先,需要提供集成验证工具来解决模式 D 用户当前通过跨多个平台的手动过程拼凑验证的实践。应规定统一的验证仪表板，提供多模型检查（在 ChatGPT、Claude、Gemini 之间运行相同查询，突出显示一致/不一致）、自动事实检查（与 Wikipedia、Wikidata、学术数据库交叉参考）、引用验证、代码执行测试以及领域专家路由。I3 的 30-45 分钟三角验证工作流程表明，实施后可将验证时间减少到 15 分钟以内。I34 的信息沙盒策略提出了额外要求：用户应能上传可信文档，创建 AI 必须优先于一般训练数据的临时可验证知识库。 其次，认识到新手用户缺乏模式 D 的怀疑立场，需要提供批判性思维脚手架来发展批判性评估习惯。应在人工智能响应后进行苏格拉底式提示（“在接受这个答案之前，问问自己：我如何验证这个？”）、针对过度自信语言的危险信号检测、没有证据的泛化识别以及因果关系声明的质疑。这些脚手架明确了专家模式 D 用户自动执行的批判性思维动作，支持新手的技能发展。 第三，需要实现透明不确定性显示，这代表最关键的要求，所有模式的 98%参与者（48/49）都提出了要求。当前的人工智能系统在模型概率中表现出校准不充分的置信度，但在流畅输出中表现出语言上的过度自信。应规定声明级别的置信度评分，为每个断言显示百分比置信度、不确定性来源归因，识别知识截止限制、模糊查询、可靠性下降的领域边界以及来源冲突。I33 的黑盒担忧表明这必须超越置信度评分到机制性解释，至少应在发生黑盒推理时标记：“此建议来自跨训练数据的模式匹配。我无法解释底层因果机制。”

在理论层面，模式 D 有三点贡献。其一，将“适当依赖”（Lee 与 See，2004）这一抽象概念进行了经验性操作化。Lee 与 See 提出用户应使信任与 AI 的实际可靠性相匹配，但对“如何做到”指导有限。模式 D 用户通过经验性测试（I42 的已知答案规程）、多源证据融贯（I22 的多模型一致）与领域阈值设定（I33 的可解释性硬性要求）提供了可操作机制，将理论落地为实践。I34 的信息沙盒增加了有界可靠性的维度：用户可以创建可靠性变得可验证的有限情境,而不是试图校准对人工智能全部能力的信任（不可知）。这表明了设计原则：系统应该使用户能够将人工智能的信息空间约束到验证可行的子集，而不是要求对人工智能的整个知识库进行信任校准。当可靠性领域有界且可测试时，适当依赖变得可实现。其二，模式 D 提出“人机同位模型”，即将 AI 视为需要审阅的同侪，而非权威或下属。不同于模式 A 中的等级定位（人高于 AI），或模

式 E 中的教学定位（人作为教师，AI 作为学生），模式 D 将 AI 概念化为一位需监督的同事，类似导师与博士后之间的关系。“AI 如同博士后”的心智模型对高专业度用户而言或许是最健康的立场，既避免将 AI 视为权威而导致的过度信任，也避免仅视其为工具而限制其潜在价值。其三，模式 D 指出在高风险领域中黑箱问题是根本性的，而非通过界面改良即可解决的。I33 对不可解释 AI 的拒绝并非技术恐惧，而是职业伦理的体现。在交易、法律、医疗等零和竞争情境中，机制透明是不可让渡的要求。以当前大语言模型为代表的不透明性在这些领域构成根本限制，提示未来的 AI 发展路径可能出现分化：在低风险应用中，允许较高不透明度以换取更强能力；在专业高风险应用中，则需以透明可解释为前提，即便能力受限。这需要探索面向专业场景的“白箱”式 AI 架构。重要局限在于，模式 D 成本高且认知负荷重，个体层面通常只能在高风险任务或高专业用户中可持续。若要实现普及，需要将模式 D“系统化”，即在系统层面内建核验基础设施（对应 MR11），使所有用户在不增加个体负担的前提下也能获得系统性检查。这意味着从“教育用户去核验”转向“把核验内嵌进系统”。

（5）模式 E：教学化反思与自我监控

在本研究样本中，模式 E 出现在 14% 的参与者身上（n＝7），是本研究所识别的最具元层面的使用模式。与以不同策略推动任务完成的 A 至 D 各模式不同，模式 E 的使用者主要将人工智能用于增强自我觉察并发展元认知能力本身。这些使用者进行的是“反思性的 AI 使用”，即把 AI 当作“镜子”，审视自身思维过程、识别认知盲点，并有意识地培养元认知技能。由此，AI 的角色从信息或生产力工具扩展为元认知发展的认知工具（Azevedo，2020；Roll 与 Wylie，2016）。

第一个特征为反向角色扮演与“AI 为学生”的模型。模式 E 的标志性行为是角色反转：要么让 AI 充当“自我”、人充当外部观察者，要么让 AI 充当“学生”、人充当教师。受访者 I5（分析学硕士生），描述了其如何使用 ChatGPT 作为模拟面试官以探查自身思维：“我把简历和研究兴趣给 ChatGPT，然后请它扮演一位‘苛刻的面试官’，专门指出我的薄弱环节。它会问：‘为何选择这种方法而不是其他备选？’‘你在做哪些假设？’‘你最大的研究缺口是什么？’这些问题迫使我把此前没有完全想清楚的内容表述出来。与单独思考相比，这有助于我更全面地反思。”当被问及这种方式与人类导师提出类似挑战性问题有何不同，I5 指出了模式 E 的独特价值主张：“人类面试官受限于其知识与偏见。GPT 借助成千上万的面试情境，可以从多个角度发问，既有哲学层面，也有方法与实践层面。这就像同时面对十位面试官。而且它不作评判，我可以承认不

确定，而不用担心形象受损。”这凸显了 AI 的一种悖论式优势：恰因为它缺乏真正的智能，使其较之人际互动更有利于诚实的自我审视；后者往往带来面子、地位与关系动态等社会成本。使用者可以承认困惑、暴露无知或探索尚未成形的想法，而无须担心被评判，从而降低了实现真实自我反思的心理门槛。

受访者 I41（理论光学博士生），提出了其称为“教学规程”的方法，将典型的 AI 辅导系统倒置：“我对 ChatGPT 说：‘你是我的学生。我会把一个物理概念分段讲解。每讲完一段，你要出题检验我是否理解了刚才讲的内容。如果我答对，我们继续；如果没有，请指出我的困惑。’这迫使我真正掌握材料，可以说是‘费曼技巧的强化版’。”所谓“费曼技巧”，即用通俗语言讲解概念以检验理解（Feynman 等，1963）。I41 的实践置于既有学习科学原则之中；而 AI 能实现人类学生难以做到的条件，包括无限耐心、能够在任意难度层级提问，并能通过与训练数据中的专家性解释进行比较来识别细微的概念漏洞。I41 指出，这一规程揭示了其此前未曾意识到的概念空白：“当我必须把量子隧穿解释得足以让‘GPT 学生’听懂时，我意识到自己在使用记忆的方程式，而非把握物理直觉。将复杂内容加以简化的行为暴露了我的困惑。”这体现了“门徒效应”（Chase 等，2009），即教学能够提升“教师”的学习。在此情境中，AI“学生”的价值不在于其是否真正“学会”，而在于它迫使使用者借由解释尝试来外化并审视自身理解的过程。

第二个特征为自我纠错回路与反思性元提示。在角色反转技术的基础上，一些模式 E 使用者发展出精巧的自我纠错机制，使 AI 成为检视其自身输出的元认知镜像。受访者 I24（软件工程研究员），描述了其称为“反思性元提示”的做法：“当 GPT 给出答案后，我不会直接接受。我会问：‘回到我最初的需求，你的答案是否需要进一步优化？’这会触发 GPT 进行自我批评，它会逐条列出我的原始需求，评估其回答对每一点的覆盖程度，识别遗漏或疏漏，然后生成改进版本。”这一技术体现了在 AI 层面的二阶元认知：使用者命令 AI 对自身输出开展元认知评估，从而形成 AI 介导的自我反思回路。I24 解释其元认知价值：“关键不在于第二个答案是否完美，而在于观察 GPT 如何批评自身，从而教会我如何批评自己的工作。我看到它会问：‘我是否回应了使用者关于可伸缩性的关切？’‘这个解释是否足够清晰？’这些也是我应当用来检视自身工程决策的问题。AI 成为我元认知过程的示范模型。”当被问及为何不直接要求“改进”时，I24 揭示了其教学意图：“如果我仅说‘改进它’，GPT 或许会把答案变好，但我学不到评估标准。让它明确说明缺了什么、为何缺失，我就能内化这套评价框架。这好比向

教师索要评分量表，立刻就能明白‘好的作品’意味着什么。”此做法与简单的输出迭代截然不同，它将元认知评估过程外显化，使原本隐性的质量判断转化为可学习的明确标准。I24 指出，这一技术源于他早期对 AI 输出不稳定的挫败感：“一开始，GPT 会给出在技术上可行但偏离要点的方案。我没有放弃 AI，而是思考如何让它‘更加慎思’。于是我开始在提示中加入‘回到……进行反思’之类的表述。这迫使 AI 和我共同后退一步，评估我们是否在解决正确的问题。”这体现了在结构层面的元认知提示，即不仅指挥 AI 产出什么，还指挥它如何评估自身的产出过程。I24 的实践在理论上的意义在于表明，AI 可以通过强制外显评价过程来充当元认知脚手架，而非依赖其更高的“智能”。通过让 AI 对自身输出“思考出声”，使用者能够洞见质量标准、缺口识别策略与改进路径，而这些往往隐含于人类专家的推理之中。这提示了设计机会：让系统自动开展自我批评，使元认知评估过程外显并可供学习。

元认知策略的外显化是模式 E 的另一项标志。受访者 I13（AI 工程师），利用 AI 对其直觉性工作流进行形式化：“我有一个模糊的‘发散—收敛’工作流，先广泛头脑风暴再逐步聚焦，但一直比较模糊。我让 GPT 帮我分析：‘我有这样一种模式：[描述工作流]。其背后的正式理论是什么？有什么局限？’GPT 将其与设计思维文献关联起来，并指出我跳过了‘共情’阶段。现在我的工作流更具原则性。”I13 还描述了其如何指挥 AI 的认知过程：“我直接告诉 GPT 应采用何种思维技术，例如‘进行逐步推理’或‘使用第一性原理分析’。我不只是给它任务，而是指挥它如何思考。这使我更加自觉地辨识可调用的不同思维模式。”这种被称为“元认知提示”的实践，是一种更为成熟的 AI 素养，使用者不仅指挥 AI 的输出，还指挥其认知过程。其类似于治疗师根据来访者需求选择认知行为治疗或心理动力学取向，只不过应用在 AI 交互上。这提示了一种新的能力层级：层级 1（任务提示：“写 X”），层级 2（策略提示：“使用 Y 方法”），层级 3（元认知提示：“按照 Z 模式思考”）。

受访者 I25（AI 与教育研究者），展示了此实践最为先进的形式，即她称为“元认知指令”的方式：“我不再用冗长提示告诉 AI 要做什么，而是命令它选择恰当的思维技术。我会说：‘分析这个问题并说明应当采用哪种推理方式，是逐步推理、第一性原理、类比推理，还是其他方法。然后据此展开。’这迫使我把注意力放在‘关于思考的思考’上，而不仅是问题本身。”将 I25 的做法与早期提示策略比较，其演进意义愈发清晰。初学者侧重任务规格（“写一份摘要”），中级使用者会加入策略引导（“用逐步推理撰写摘要”），而 I25 在元层面运作，她把策略选择本身委托给 AI，同时保持评价性

监督：“GPT 可能会说：‘对于该问题，建议采用比较分析，因为你在评估多个选项。我将按照标准识别、逐项评估与综合给出结构。’看到这样的结构，我会判断：比较分析是否确为恰当心智模型。有时我会认为不然，该问题更适合系统思维。无论哪种情形，我都在发展自己的策略谱系。”这代表了最高层级的认知过程外显化，不仅是执行元认知策略，而是构建关于“有哪些策略、何时适用、如何抉择”的元认知意识。I25 解释其学习机制：“每当 GPT 说明它为何选择某种思维技巧时，我就在累积自己的认知策略‘库’。这好比观看名厨解释为何使用某把刀或某种烹饪法，你不仅学到技巧，还理解了背后选择技巧的决策过程。”这里的理论跃迁在于从“借助 AI 思考”转向“借助 AI 学会如何进行关于思考的思考”。I25 的实践表明，AI 对元认知发展的价值不主要在于输出结果，而在于其能够使专家级的认知策略选择可见、明确并因此可学。这与认知学徒制理论相关（Collins 等，1991），在该理论中，专家通过示范与表述使隐性知识可见；此处，AI 充当耐心的“认知导师”，展示并解释其策略选择过程。当被问及成效时，I25 指出了元层面的收益：“我的问题求解更加有意图。过去我往往凭直觉下手，现在我会停一步思考：‘这里恰当的认知取径是什么？’有时我仍然使用 AI，但越来越多时候我在独立运用这些策略。我的目标从来不是依赖 AI，而是发展自身的策略性思维能力。”这正是模式 E 的发展性意识：AI 是构建持久元认知能力的临时脚手架，而非永久性的认知假肢。

技能退化的觉察与主动性对策是模式 E 的第四个特征，与无意识过度依赖的模式 F 不同，模式 E 的使用者会主动监测自身依赖并采取保护性措施。受访者 I38，数据科学本科生，描述了其从无意识过度依赖到有意识技能保护的历程：“使用的第一个月，我的信任峰值达到 90%，我觉得 GPT 很神奇。但我注意到自己在写作中越来越难以表达本意。我不得不‘和 GPT 暂时分手’。现在我的信任稳定在 20%。我让它处理大量低层次思维任务，但把所有高层次推理完全留给自己。”I38 的觉醒时刻发生在真实绩效要求情境中：“我尝试不依赖 GPT 写一封邮件，却发现难以找到恰当的措辞，这些词原本是我熟悉的。这让我警觉。输出速度不等于真实能力。所谓生产率提升是一种错觉，就像兔子坐车战胜乌龟，看似更快到达，但并未训练出奔跑能力。”这一隐喻呈现了模式 E 的发展性意识：过程具有超越结果的内在价值，认知努力本身不仅服务于结果，也关乎学习与技能维持。来自预访谈而未纳入主分析的受访者提供了最为极端的主动技能保护个案：“我在工作中大量使用 ChatGPT，但在个人项目里刻意降低其角色。我仍然在副业项目中完全自行编码以维持技能。这就像运动员在训练时不用器械，需

要保留原生能力。”这种“刻意低效”，即在存在更快的 AI 辅助方法时选择较慢的手动路径，体现了对技能维护的有意识投入，承认能力若无定期训练便会萎缩。体育训练的隐喻提示，模式 E 的使用者将认知技能视为与体能相仿，需要维护与锻炼。

过度依赖的心理维度在模式 E 的叙述中也十分明显。I38 描述了信心下降：“当我离开 AI 就写不出来时，我觉得自己很‘笨’，这很尴尬。”受访者 I47 则表达了焦虑：“我害怕失去英语写作能力。”这些情绪反应，包括羞耻、焦虑与自我效能感下降，说明 AI 过度依赖不仅威胁工具性能力，也威胁使用者的身份与胜任感。这与自我决定理论相一致（Deci 与 Ryan，2000）：个体需要自主（对工作的控制）、胜任（能力感）与连结（与他人的关系）。过度委托 AI 会削弱自主与胜任，即便在 AI 协助下客观任务表现尚可，个体仍可能出现心理困扰。

对于模式 E 元认知过程， 在本数据集中呈现出最高的总体元认知复杂度，评分为 8.6/12（SD＝1.6），其中评价分量表的表现尤为突出，得分为 9.4/10。规划过程的证据强度为 78%，但与模式 A 的任务分解式规划在性质上不同，模式 E 的规划更加关注自我知识目标，而不仅是任务完成，并且区分探索性学习（为形成理解而尝试新方法）与利用性执行（以高效方式使用已验证方法）。监控过程的证据强度为 58%，包括对技能退化的觉察与依赖度跟踪，但不及模式 D 在输出核验方面的强度。评价过程的证据强度为 94%，表现为深度自我反思、元层面批评（例如“我是在学习，还是仅在完成任务”）以及对过程价值的承认。调节过程的证据强度为 72%，体现在有意的策略调整、脚手架淡出（例如 I38 的“分手”代表主动降低 AI 依赖）与角色操控（例如 I41 在“教师—学生”角色之间切换）。区别模式 E 与其他有效模式的关键在于二阶元认知，即“关于思考的思考”。模式 A 体现一阶元认知（管理任务过程），模式 D 强调一阶监控（验证输出），而模式 E 在递归的元层面上审视元认知过程本身。I13 询问 AI“何种正式理论解释我的工作流”的实践即为例证，他不仅在使用元认知策略（一阶），而且在研究与改进这些策略（二阶）。类似地，I24 的自我纠错回路与 I25 的元认知指令均属二阶操作，使用者在其中检视并优化自身的元认知路径，而不仅仅是应用这些路径。

量化分析表明，模式 E 是第三常见的模式，占 16%（8/49）。在主观学习收益方面，86% 的模式 E 使用者报告因使用 AI 而“理解更深入”，而模式 F 使用者仅为 34%（χ^2＝12.7，p＜.001），表明模式 E 的方法确能转化为更佳的主观学习体验。策略觉察达到 100%，即所有模式 E 使用者均能清楚表述其元认知策略，而总体样本仅为 44%（21/49）。技能保护行为出现在 86%（6/7） 的模式 E 使用者中，他们会主动监测依

赖程度，而总体样本为 21%（10/49），进一步印证了模式 E 在技能退化风险意识方面的独特性。出现一项独特发现：模式 E 与教学经验呈正相关（r＝0.58，p＝0.001）。曾有教学他人经验者，无论是助教、导师还是职业教育者，更可能呈现模式 E。这表明教学情境中的元认知可迁移至 AI 使用。教学要求表述默会知识、识别学生误解、监测理解并调整解释方式，而这些元认知技能在 AI 场景中同样适用。教师因专业角色所需而自然进行“关于思考的思考”，从而更易将 AI 视为元认知发展的媒介，而不仅是生产力工具。

模式 E 启示了四项支持元认知发展的设计需求。首先，需要建立引导反思机制,将苏格拉底式提问嵌入 AI 交互。系统不应仅提供答案，还应引导反思，例如：“继续之前，请先思考：你理解的是其为何有效，还是仅知道其有效？”具体实现包括理解关卡（参照 I41 的“学生模型”），要求使用者在继续之前进行展示。基于 I24 的反思性元提示，系统应自动触发自我批评机制，将 AI 由答复者转变为元认知教练。 其次，认识到模式 E 使用者常通过试错发现有效策略,系统需要提供元认知策略指导。实现方式包括入门教程，呈现“专家使用者的有效 AI 使用策略”，例如发散—收敛模型、“从 0 到 1”原则、教学规程、自我纠错回路等。还可建设策略库，允许使用者贡献与投票。对于高阶使用者，系统应提供“元认知提示模式”，明确引导认知策略选择，将 I25 的实践操作化。 第三，回应 I38 的经验,需要实现技能退化预防机制。系统应持续监测独立完成与 AI 协助的比例，提供使用分析看板显示“独立工作：32%（此前为 45%）”。当独立性持续低于阈值时，系统应发出提醒。使用者可设定“无 AI 日”，系统在该时段屏蔽 AI 以强制技能维护。还可引入游戏化元素以鼓励独立性。 第四，应为模式 E 使用者提供学习过程可视化,展示任务复杂度随时间的变化、关键突破时刻、“有无 AI”表现对比以及元认知里程碑。可构建策略演化时间线,显示使用模式的发展轨迹。该设计契合模式 E 的反思取向，使学习过程可见并强化成长。

模式 E 具有四方面理论贡献。第一，它通过引入 AI 介导的反思，扩展了 Schön（1983）的“反思型实践者”理论。Schön 指出，专业人士通过对行动的反思从经验中提炼教训。模式 E 表明，AI 能充当“第二自我”，使个体得以与自身思维对话，从而产生仅靠内部反思难以达成的元层面加工。这不同于写日记（独白）或导师指导（与他人对话），因为 AI 能作为可控、耐心且具多元视角的对话伙伴，像 I5 所言“仿佛同时面对十位面试官”，从多个角度再现使用者的思考，使个体能够对自身认知进行换位审视。I24 的自我纠错回路进一步表明，AI 可以示范元认知评估过程，使隐性的质量标准变

得明确且可学。第二，模式 E 颠倒了智能辅导系统中的传统教学代理模型。常规智能辅导系统通常设定 AI 为教师、人为学生（Koedinger 与 Corbett，2006）。模式 E 展示了反向配置的价值，即人为教师、AI 为学生。这与“通过教学进行学习”的文献一致（Roscoe 与 Chi，2007），但新增了一个要素，即 AI 的无限耐心与广泛覆盖，使人类讲解的深度超越一般学生的理解力与注意力所限。设计启示是，AI 辅导系统应提供双向角色，允许使用者根据学习目标在被 AI 教授与向 AI 教授之间切换。第三，模式 E 将元认知提示界定为高级的 AI 素养，并提供层级化成熟度的证据。I13 指挥 AI 采用“逐步推理”等思维技术的实践，质上不同于一般的任务提示。I25 的元认知指令，即在保持评价性监督的同时，将“认知策略选择”委托给 AI，代表更高层次。这提示了一种更为精细的能力框架：层级 1（任务提示：“写 X”），层级 2（策略提示：“使用 Y 方法”），层级 3（元认知提示：“按照 Z 模式思考”），层级 4（更高阶元认知提示：“决定应采用哪种思维方式并给出理由，然后据此展开”）。层级 4 需要对 AI 能力与通用认知策略的广泛理解，以及评价策略适切性的能力，代表了 AI 素养的前沿，值得后续研究重点关注。第四，模式 E 揭示了教育类 AI 讨论中的效率与学习之间的张力。I38 的“兔子与乌龟”的隐喻以及 I41 在订阅短暂中断后“重新享受编码”的叙述均表明，过程具有超越结果的内在价值。自我决定理论（Deci 与 Ryan，2000）可解释这一点：人需要自主、胜任与连结。过度委托会削弱自主与胜任，导致自我效能感下降。I38 所言“当我过度依赖 AI 时，我感觉自己不再像一个真正的数据科学家”，正是因过度依赖而产生的身份威胁。关键洞见在于，AI 采用应优化个体福祉，而非仅仅追求生产率。在某些使用者与情境中，即便短期产出效率降低，较慢的手工路径在心理上更为健康。

（6）模式 F：无效或被动使用与过度依赖

尽管本研究主要关注能够促进学习的有效使用模式，但理解无效模式对于设计干预同样至关重要。模式 F 代表元认知参与的缺失，而非存在。值得注意的是，在本研究的 49 名直接访谈对象中，没有受访者将模式 F 作为其主要使用模式——这本身反映了自我选择偏差，即愿意深度讨论“AI 使用经验”的参与者往往已具备一定元认知觉察。 因此，对模式 F 的理解主要源自三个来源：（1）12 位教师/导师关于其学生的二手报告，（2）少数受访者（如 I38）对其早期"觉醒前"阶段的回顾性叙述，（3）教师在 200 余名学生课堂中的系统观察。其特征包括对 AI 输出的不加分辨的接受、缺乏核验行为以及无意识的技能退化，即使用者将 AI 视为"魔术"，而非需要批判性参与的工具。

第一个特征为任务完全外包与核验缺失。模式 F 的典型行为是完全外包任务，不进行分解与监督。受访者 I24（软件工程研究员），描述了这一模式：“一些学生提交我称之为‘GPT 口味’的论文，让人一眼就能看出来。语言很润色，但逻辑很糟。当我要求他们解释推理过程时，他们会很吃力。他们外包的不仅是写作，连思考也外包了。最可怕的是，他们没有意识到逻辑缺陷，因为他们从未与思想进行深度交互。”这种表层流畅与深层理解之间的脱节，正是模式 F 的核心问题：AI 的语言流利性掩盖了概念层面的缺陷，从而制造“能力”的错觉。I24 总结了三种预警信号，以区分模式 F 的作业与合理使用 AI 的作业：其一是过度自信的流利性（语法完美却缺乏实质内容），其二是脆弱的理解力（无法回答关于自身作品的基本澄清问题），其三是防御性的转移（当受到质询时，以“研究就是这么说的”回应，却说不清具体是哪项研究以及为何如此）。这些迹象显示出被动的内容消费，而非对 AI 生成内容的主动整合。

受访教师 I10（金融学副教授），提供了一个展示模式 F 风险的警示案例：“一位学生使用 GPT 生成的期权定价代码。公式看上去正确，记号规范，推导整洁，但在边界条件上存在一个细微错误，在真实交易中会造成灾难性损失。当我指出这一点时，该学生回答‘但 GPT 通常是正确的’。问题在于，‘通常’在金融领域是不够的。”此事表明，模式 F 的盲目信任会将小概率的 AI 错误转化为高后果的现实失败。学生诉诸 AI 的一般性可靠性，而非与特定错误交涉，这恰是模式 F 的认识立场：将 AI 当作无误权威。受访者 I48（电气工程本科生），罕见地提供了模式 F 的直接自述：“我用 AI 帮我操作诸如 Aspen Plus 之类的复杂软件。我很少核验，因为它很技术性，我假设它比我更懂。有时它会犯基础计算错误，例如把 5×2 算成 11，或者给出错误的变量值，但我通常只在最终答案显得明显不合理时才会发现。”当被问及为何不系统核验时，I48 揭示了根本性误解：“核验要花更多时间，不如我自己重做。用 AI 的意义在于速度。如果我必须检查一切，那还有什么好处？”这种框架把核验视为否定 AI 价值，而不是有效使用 AI 的有机组成部分，这一误解也受到 AI 市场宣传中“轻松与快速”的强调所强化，而这些宣传很少附带关于准确性要求的警示。

第二个特征是无意识的技能萎缩与防御性合理化。I38 在前述模式 E 中为技能退化机制提供了回溯性洞见：“使用 ChatGPT 的第一个学期，我的成绩提高了，论文更为精致，代码更为整洁。但我注意到一件可怕的事：我在表达自己想法的能力方面出现了退化。我坐下来写作，却找不到曾经熟悉的措辞。想法仍在，但我把表述外包给 AI 的时间太久，导致表述能力已经萎缩。”这种渐进而无意识的退化是模式 F 的隐蔽

危险所在：产出质量的提升掩盖了底层能力的衰退。觉醒时刻发生在要求“去中介化”的绩效情境中：“面试官让我不看笔记来描述我的项目。我表现很糟，措辞含混，无法构建清晰结构。我已让 GPT 代我构建结构数月，不仅写作退化，思维也退化了。”来自预访谈的 受访者也 以生动隐喻表述：“我感觉自己正在失去‘认知肌肉张力’。所有事情都由 AI 中介，甚至给朋友写随意邮件也要用它。有一天我尝试手写简单邮件，却花了 20 分钟。我想，这太离谱了，我以前可是一个写作不错的人。于是我果断断用两个月以重置状态。”认知肌肉萎缩的隐喻强调，技能需要规律训练，而非偶尔为之。观看 AI 写作而只提供高层级指令，就像看健身视频却久坐不动，旁观表现并不等于能力增长。

当被指出过度依赖时，模式 F 使用者往往采取防御性合理化，而非调整策略。常见的合理化包括：“大家都在用”（社会印证，即若广泛使用便是可接受）、“教授不理解现代工具”（权威贬损，即批评者与时俱进不足）、“我通过观摩 AI 的好例子在学习”（学习谬误，即仅观摩范例而不进行产出性练习，不会形成程序性知识）、以及“重要的是我的想法，而非写作本身”（错误的内容—形式割裂，写作本身就是思考，而非单纯的文字转录）。I24 指出，这些合理化是防御性的，而非反思性的：“学生们在保护自我，而不是直面关于依赖的令人不适的事实。当我指出他们无法解释自己的工作时，他们会自我防御，而不是开始担忧。”

在元认知缺位方面，模式 F 在所有维度上都表现为极低的元认知。规划过程的证据仅为 15%，目标集中在“快速完成”而非“学习”或“维持技能”。监控过程几乎缺失，仅为 8%，没有过程跟踪与理解检验，错误往往只在外部权威（导师、主管）指出时才暴露。评价过程的证据仅为 12%，倾向将 AI 视为无误而不进行能力判断，使用者往往要到危机时刻才意识到自身技能退化。调节过程几乎缺席，仅为 5%，使用者很少进行策略调整，无效模式会持续存在，直到不良成绩、公开场合的尴尬失败或直接干预迫使改变。刻画模式 F 的关键元认知失败是“元认知无觉察”，即“不知道自己不知道什么”，这与著名的达宁—克鲁格效应相呼应（Kruger 与 Dunning，1999）。AI 的流利输出制造了“流利幻觉”，让使用者因作品“看起来专业”而感觉自己“知识充足”，将 AI 的表现与自身理解混为一谈。这是“代理能力”的一种形式，即因 AI 能够产出高质量结果而误以为自身也具备相应能力，尽管使用者既不能独立产出，也不能有效评估这些产出。

由于自我报告偏差（呈现模式 F 的使用者不太可能主动参与“有效 AI 使用”研究）、对二手教师报告的依赖及其过渡性（许多使用者在危机触发调整前会暂时呈现模式 F），模式 F 的确切流行度难以界定。基于三位教师在 200 余名学生的课堂中的观察所做的保守估计显示，25% 至 40% 的学生呈现模式 F 行为，但鉴于教师更易注意到问题个案，这一比例很可能是上限。直接观察中仅有 8%（4/49）承认在“觉醒”前呈现过模式 F，但这几乎可以肯定低估了真实比例。教师与受访者访谈的后果主要集中在三类领域。学业后果包括较低的理解水平（I24 的课程数据显示，87% 的模式 F 学生在口试中表现不佳，而已采用核验策略的学生为 23%）、在 AI 不可用的时间压力下无法表现，以及因被教师识别而引发的学术诚信调查。职业后果包括面试中的尴尬（I38 的失利）、主管的不信任（I10 在期权定价事件后不再允许团队成员不受限地使用 AI），以及在 AI 系统失效或不可用时无法胜任工作。心理后果包括信心下降（I38：“当我离开 AI 就写不出来时，我觉得自己很笨”）、依赖焦虑（I47：“我害怕失去英语写作能力”）与身份威胁（I38：“当我过度依赖 AI 时，我感觉自己不再像一个真正的数据科学家”）。

模式 F 是一个需要通过有效模式来解决的设计问题。问题不在于“如何支持模式 F 使用者”，而在于“如何防止模式 F 的出现，并引导使用者向有效模式（A—E）转变”。这需要承认某些看似理性的短期优化行为（最小化努力、最大化成绩）会带来有害的长期后果（技能萎缩、学习失败），并据此采取干预性的设计理念。模式 F 提示了三项设计需求，这些需求旨在预防模式 F 的出现并引导使用者向有效模式转变。 首先，需要建立过度依赖检测与干预机制。系统应监测模式 F 指标，包括用户提交超过 500 字的提示（任务完全外包）、未观察到核验行为、在 30 秒内即接受输出,以及跨多次会话的稳定化模式。一旦检测到，应实施干预：“学习成效提示：你最近连续 8 次接受输出时未进行核验或测试。研究显示，此模式通常导致浅层理解与长期技能退化。”对于持续呈现模式 F 的使用者，可引入“强制摩擦”机制，例如在允许复制之前要求使用者先用自己的话总结 AI 输出。其次，针对模式 F 的核心问题，即使用者既缺乏对元认知策略必要性的觉察，也缺乏对这些策略内容的了解，需要提供渐进式元认知脚手架。第一阶段（觉察建构）通过教育性插页提示“专家会如何核验”、社会比较信息提升意识。第二阶段（技能发展）提供在引导下的实操,配有检查清单与自适应提示。第三阶段（独立性）逐步淡出脚手架,辅以抽查与正向反馈。 第三，认识到模式 F 在某些情境下可能反映“理性的优化”（如低风险、高时间压力任务），系统需要支持学习模式与

生产模式的切换。学习模式提供高强度脚手架、强制反思、必需核验与较慢节奏,适用于新技能学习。生产模式提供低脚手架、快速输出、自选核验，适用于常规任务。关键的安全装置在于，当系统识别到用户在看似高风险任务上选择生产模式时，应给出提示，使模式选择成为有意识的决定而非不加反思的默认。

模式 F 揭示了人—AI 协作中的三项理论张力。第一，它指向元认知悖论：最需要元认知支持的人往往最不可能意识到这种需要。这带来设计悖论，即最能从 MCA 的元认知功能中受益的使用者最不可能主动使用这些功能。解决之道在于将元认知功能设为“默认开启而非自愿选择”，通过智能默认与适度的温和家长式干预来保护新手，这对强调使用者自主性的传统以使用者为中心的设计提出挑战。第二，它呈现了所谓“流利幻觉”：AI 的语言流畅性制造“代理能力”，让使用者因输出“看上去专业”而感觉自己“知识充足”，将 AI 的表现与自身理解混淆。这代表了达宁—克鲁格效应的一种新形式，不再是自我评估的失准，而是被 AI 的能力所遮蔽。理论上的含义在于，人—AI 系统需要开发区分“学习”与“表现”的新型评估指标。学生或许可以凭借 AI 辅助在作文上获得高分（高表现），但在学习上毫无收获（零迁移）。传统只测结果的教育评估无法捕捉这种分离。后续研究需开发与表现相独立的学习评估方法，例如迁移测试、解释性要求或在 AI 不可用条件下的表现测评。第三，它揭示了效率与学习之间的根本权衡：AI 的核心价值主张在于节省时间，而学习的要求恰在于投入时间并进行富有成效的“挣扎”。对于学生与学习者而言，优先效率可能导致灾难性结果，即快速完成作业而毫无学习。设计应帮助使用者意识到速度并非总是正确目标。系统若以效率为最优先，便会不经意地鼓励模式 F。以学习为中心的系统必须有意牺牲部分效率以促进元认知，这与当前强调“零摩擦、对话顺畅、即时结果”的 AI 设计理念形成了根本反转。在 MCA 的设计理念中，只要能促进元认知、从而促进学习，“摩擦”就是特性，而非缺陷。

3.4.3 跨模式分析：元认知维度与有效性

在详细描述了六种模式之后，本节从多个维度考察有效使用与无效使用的区别因素，回应我们的核心问题：为何背景相似的用户（研究生、计算机科学专业、频繁 AI 使用者）会表现出截然不同的模式？我们的分析揭示，答案在于元认知参与而非人口统计学特征。

（1）元认知复杂度：量化对比

我们基于观察到的独特元认知子过程数量计算元认知复杂度分数（0-12 分），每个子过程根据证据强度评分为强（✓✓✓）、中（✓✓）或弱（✓）。12 个子过程跨越四个类别：规划包括 4 个子过程（任务分解、目标设定、策略选择、角色定义）；监控包括 3 个子过程（过程追踪、质量检查、信任校准）；评估包括 3 个子过程（输出质量评估、风险评估、能力判断）；调节包括 2 个子过程（策略调整、工具切换）。如果子过程在用户访谈数据中显示中等或强证据（✓✓或✓✓✓），则计入复杂度分数。

图 3-2 展示了各模式的元认知复杂度分布，揭示了有效模式（A、B、C、D、E）之间的显著差异。模式 A 平均 8.1/12（SD=1.1，n=18），模式 C 为 6.4/12（SD=1.3，n=16），模式 E 为 8.6/12（SD=1.6，n=7），模式 D 为 8.5/12（SD=1.2，n=4），模式 B 为 6.0/12（SD=1.4，n=4）。总体样本均值为 7.5/12（SD=2.3，N=49）。单因素方差分析证实模式间存在显著差异（$F(4,44)=38.4$，$p<.001$）。事后 Tukey 检验显示，模式 E 和 D 得分最高但彼此无显著差异（$p>.05$），模式 A 得分亦处于高位且与 E、D 无显著差异，模式 B 和 C 显示中等分数且彼此无显著差异（$p=.67$）。

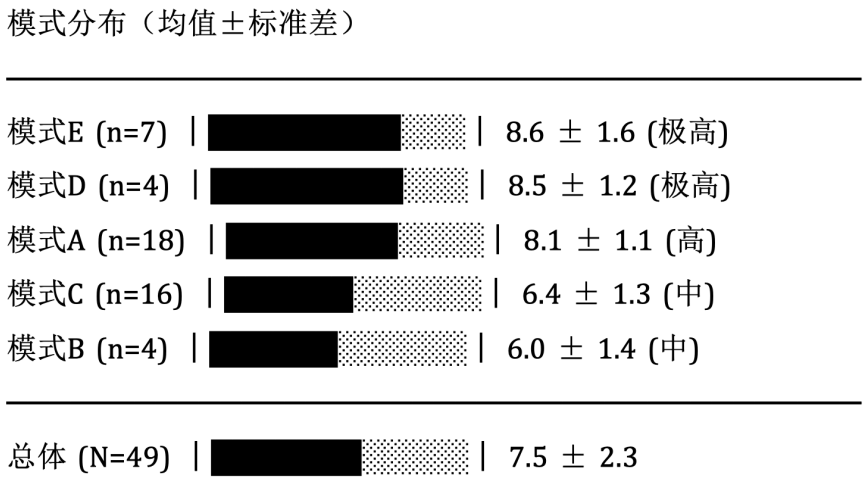

图 3-2 各模式的元认知复杂度分布

资料来源：本文研究

这一量化对比的关键洞察在于，虽然有效模式在总体元认知复杂度上相近（均≥6.0/12），但它们在四个元认知 维度上表现出不同的侧重模式。图 3-3 展示了五种有效模式在规划、监控、评估、 调节四个维度上的得分画像（满分 10 分）。

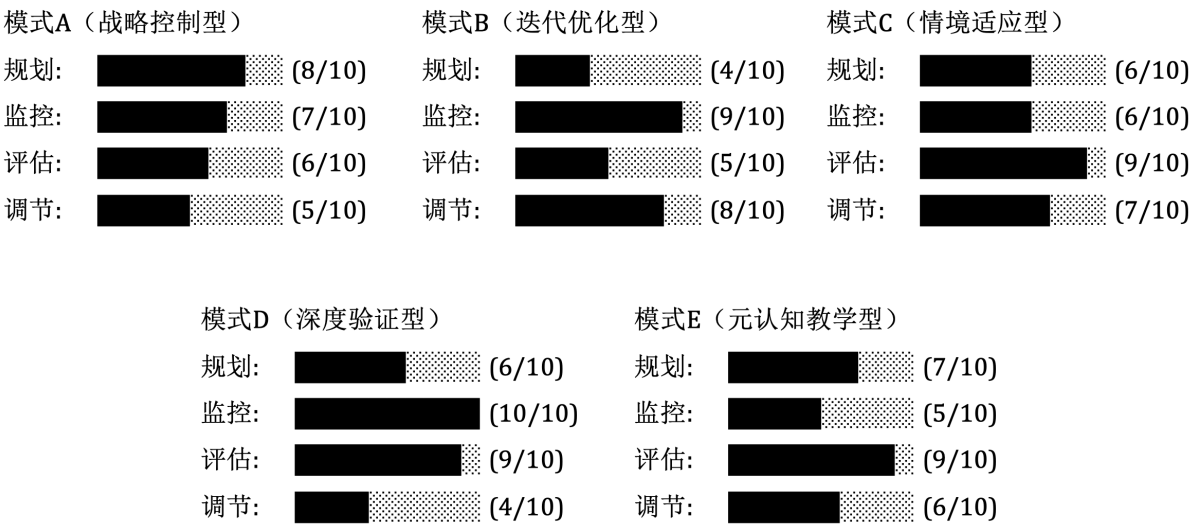

图 3-3 模式的元认知画像对比图

模式 A（战略控制型）在规划（8/10）和监控（7/10）维度表现突出，体现其“前瞻性控制”特征：通过周密规划和持续监控维持对 AI 交互的掌控。模式 C （情境适应型）在监控（9/10）和调节（7/10）维度得分最高，反映其根据 情境动态调整策略的灵活性。模式 D（深度验证型）的监控维度达到满分（10/10）， 评估维度也很高（9/10），印证其对输出质量的极致追求。模式 E（元认知教学型） 在评估（9/10）和规划（7/10）维度表现出色，体现其对 AI 交互过程的深度反思。 模式 B（迭代优化型）展现相对均衡的画像，在调节（8/10）维度略高于其他。

这些差异化的元认知画像表明，高元认知复杂度可以通过不同的维度组合实现：模式 A 通过“强规划+强监控”，模式 C 通过“强监控+强调节”，模式 D 通过 “极致监控+深度评估”，模式 E 通过“强评估+系统反思”。这一发现支持了 “多路径有效性”理论：只要维持足够的元认知投入（总分≥6.0），用户可以 根据个人偏好和任务特征选择不同的元认知策略组合。设计启示是 MCA 系统应 支持多样化的元认知路径，而非强制所有用户采用单一“最佳实践”。

（2）有效模式的共同特征（A-E）

尽管具体策略和优先级存在表面差异，五种有效模式在三个基本特征上具有共性，这些特征将它们与模式 F 区分开来。

特征 1：主动的元认知监控

所有有效用户在三个维度上保持持续觉察：自我知识（理解自己知道什么与不知道什么，称为认识论觉察）、AI 能力（理解 AI 能够可靠完成什么与不能完成什么，

称为工具觉察）以及错误可能性（识别错误可能发生的位置，称为风险觉察）。表 3-6 展示了各模式中监控行为的流行率，揭示了有效模式与模式 F 之间的显著差异。

表 3-6： 监控行为流行率

行为	模式 A	模式 C	模式 D	模式 E	模式 B
追踪变化/输出	95%	63%	100%	67%	100%
接受前验证	89%	38%	100%	56%	95%
动态校准信任	68%	94%	75%	78%	50%
检查自身理解	84%	44%	75%	100%	40%
平均监控强度	84%	60%	88%	75%	71%

监控强度与整体有效性呈强相关（r=.76，p<.001）。有效模式的用户平均将 23% 的交互时间用于监控活动（追踪变化、核实事实、测试输出、反思理解），而模式 F 用户仅为 2%（t=11.4，p<.001）。这种监控时间投入上的 10 倍差异代表了有效与无效 AI 使用之间最清晰的行为区分标志。关键在于，监控不是被动观察，而是主动的认知工作——将输出与期望进行比较、检查内部一致性、测试论断、评估合理性。模式 F 用户跳过这一工作，将 AI 输出视为最终答案而非需要验证的中间产品。

特征 2：战略性而非机会主义的 AI 使用

有效用户对何时以及如何使用 AI 有明确理由，能够阐明从正式算法（I4 的投资回报率计算）到直觉原则（I3 的“0 到 1 规则”）的决策逻辑。相比之下，模式 F 用户使用 AI 仅仅因为“它可用”和“它更快”，而不对 AI 辅助是否符合学习目标或能力发展需求进行战略考虑。

比较决策逻辑可以说明这一区别。模式 A 用户遵循的逻辑是：若（我能以合理努力自行完成任务）则（手工完成或仅使用 AI 进行低级辅助）否则若（任务需要不熟悉的领域知识）则（使用 AI 进行 0→1 启动，然后人工精炼）否则（不使用 AI）。模式 C 用户实施：任务价值=函数（重要性、利害关系、可验证性、错误成本）；预期 AI 质量=函数（任务类型、领域、复杂度）；投资回报率=节省时间÷（修订时间+风险成本）；若（任务价值=高且预期 AI 质量=低）则（手工）否则若（投资回报率>阈值）则（使用 AI 并配以适当验证）否则（手工）。模式 F 用户遵循：若（AI 可用且任务需要努力）则（使用 AI）——没有进一步条件、没有验证、没有战略考虑。

这一差异反映了根本取向：有效用户将 AI 视为战略性部署的工具，服务于包括学习、技能维护和能力发展在内的目标；模式 F 用户将 AI 视为节省劳力的设备，在不考虑发展成本的情况下最小化努力。前者实施手段-目的分析（Newell & Simon, 1972），

评估 AI 使用是否服务于更高层级目标；后者表现出行为经济学所称的"现时偏好"（O'Donoghue & Rabin, 1999）——优先考虑即时努力最小化而非长期能力发展。

特征 3：持续学习与适应

有效用户的策略基于经验而演化，展现出我们称为"策略成熟"的特征，即通过反馈驱动的精炼。模式 F 用户尽管积累了失败经验，却停留在初始方法，不随时间调整。受访者 I16 的演化体现了模式 A 用户的策略成熟特征。第一阶段（第 1-2 个月）：AI 辅助人类："我写代码，GPT 填充函数。"第二阶段（第 3-6 个月）：共同责任："我在注释中写规格，GPT 编写 50%的代码。"第三阶段（第 7-12 个月）：角色反转："我管理项目，GPT 是程序员（80%代码）。"第四阶段（第 13 个月+）：成熟协作："我使用双 AI 系统（ChatGPT+Gemini）并刻意进行线下练习以维持技能。"这一进展展示了日益增长的复杂性：委托增加但仍受技能维护意识约束，验证系统变得更加精细（单 AI 到双 AI），并出现关于过度依赖风险的元觉察，促使主动采取对策。受访者 I4 的信任校准轨迹体现了模式 C 的演化。初始方法（第 1 个月）：对所有情境一律 80%信任。发现学术错误后进行调整（第 3 个月）：领域分化，日常任务 80%信任但学术工作 0%信任。精炼理解（第 10 个月）：形成多因子模型："重要性×预期质量×修订成本"，按任务动态应用。这一演化反映了从失败中学习（学术错误触发校准）、日益复杂（从二元到连续信任）以及原则提取（从隐性直觉发展出显性公式）。

与模式 F 形成对比：用户没有表现出战略演化，尽管积累了错误和挫折，第 1 个月和第 12 个月应用相同方法。当被问及为何在失败后不调整策略时，典型回应包括："我觉得只是运气不好"（将失败归因于随机性而非系统性问题）、"我需要写更好的提示"（将失败归因于执行而非根本方法）或"我不知道还有别的方法"（缺乏对替代策略的接触）。这种停滞反映了元认知评估的缺失：失败没有被分析以获取教训，经验中没有提取模式，也没有考虑策略调整。

（3） 元认知觉察连续体

用户存在于元认知觉察的连续体上，跨越五个层级，而非二元的有效/无效分类。在最低层级（模式 F），用户对自身知识缺口没有觉察，不加质疑地接受输出，将失败归因于外部（"AI 错误"、"运气不好"），尽管反复失败也不进行策略调整。在萌芽层级（模式 B），用户通过试错识别错误，针对失败迭代调整提示，但战略规划有限，学习是反应性而非主动性的。在情境层级（模式 C），用户认识到情境重要性并按领域和利害关系调整信任，但觉察仍局限于即时情境，缺乏系统性原则。在系统层级

（模式 A 和 D），用户实施明确的过程控制，维持原则性边界，实施全面验证，并显示对包括技能退化风险在内的长期影响的觉察。在反思层级（模式 E），用户展示对自身思维的元觉察，使用 AI 增强自我知识，监控其元认知本身，并将 AI 使用理解为发展过程。

这一连续体揭示了层级之间的移动并非随经验自动发生。一些用户在定期使用 AI 12 个月以上后仍停留在最低层级（模式 F 持续），而另一些则在数周内达到反思层级（快速出现模式 E）。这提出了关键问题：什么预测了沿连续体的进展？我们进行了多元线性回归分析（N=49），以元认知复杂度为因变量。自变量包括人口统计学变量（教育水平、学科、年龄）和经验性预测因子（教学经验、元认知训练、重大 AI 错误、AI 使用频率）。表 3-7 呈现完整回归结果。

表 3-7： 元认知复杂度的回归分析 (N=49)

预测变量	非标准化回归系数（B）	标准误（SE）	标准化回归系数（β）	t 值	显著性水平（p 值）
人口统计学变量					
教育水平（研究生）	-0.28	0.57	-0.14	-0.50	.618
学科（STEM）	0.27	0.39	0.13	0.70	.491
年龄	-0.04	0.04	-0.22	-1.13	.266
经验性预测因子					
教学经验	0.96	0.60	0.47	1.59	.119
元认知训练	1.71***	0.43	0.69	3.98	<.001
重大 AI 错误	-0.12	0.43	-0.05	-0.29	.774
AI 使用频率	0.22	0.17	0.23	1.27	.212

注：$R^2 = .507$, 调整后 $R^2 = .422$, $F(7, 41) = 6.02$, $p < .001$. ***$p < .001$.

结果揭示了显著的模式：所有人口统计学变量均无显著效应：教育水平（$\beta = -0.14$, $SE = 0.57$, $t = -0.50$, $p = .618$）、学科背景（$\beta = 0.13$, $SE = 0.39$, $t = 0.70$, $p = .491$）、年龄（$\beta = -0.22$, $SE = 0.04$, $t = -1.13$, $p = .266$）。这一发现根本性地挑战了基于人口统计学特征的预测：研究生并不比本科生更具元认知性，理工与非理工学科无差异，年长者也未自然表现出更好的实践。相比之下，经验性预测因子显示出强有力的效应模式。元认知训练成为最显著的预测因子（$\beta = 0.69$, $SE = 0.43$, $t = 3.98$, $p < .001$），接受过明确元认知策略指导（如何规划、监控、评估、调节其学习）的用户比未接受训练者的元认知复杂度平均高 1.71 分（约 23%）。这表明元认知是可学习的技能而非固定特质，可以通过正式课程、导师指导或结构化干预有效传授。受访者中，接受过此类

训练的 10 人（20.4%）平均元认知复杂度达 9.20/12，而未接受训练的 39 人仅为 7.11/12，差异达 2.09 分。教学经验显示出可观但边际显著的效应（$\beta = 0.47$, SE = 0.60, $t = 1.59$, $p = .119$）。曾教过他人的 20 位参与者（40.8%），无论是正式教学（副教授、研究员）、担任研究生助教，还是非正式辅导，平均元认知复杂度为 8.20/12，比无教学经验者（7.00/12）高 1.20 分。虽未达到传统显著性阈值，但效应量可观且方向与理论预期一致。教学要求将隐性知识外显化、监控他人理解、调整教学策略，这些元认知技能自然迁移到 AI 使用中的自我监控和策略调节。

然而,两个预期的预测因子在本研究中未显示显著效应。重大 AI 错误的效应几乎为零（$\beta = -0.05$, SE = 0.43, $t = -0.29$, $p = .774$），与理论预期的“警醒效应”不符。9 位被识别为经历过重大 AI 错误的参与者（18.4%），如 I38 的面试失败、I17 的财务损失，以及高风险领域（金融、医疗）的潜在错误，其元认知复杂度（M = 7.22）与其他参与者无显著差异。这可能反映了识别标准的局限性：“重大错误”能需要更严格的操作化定义,特指那些造成实质性损失并触发深刻反思的“临界事件”。仅有失败经历不足以促进元认知发展，关键在于失败后是否进行系统性反思和策略调整。部分经历错误的用户可能采取回避或外部归因，未能从经验中学习。AI 使用频率同样未显示显著效应（$\beta = 0.23$, SE = 0.17, $t = 1.27$, $p = .212$），且方向为正向而非预期的负向关系。本研究中 AI 使用频率（M = 4.96 天/周, SD = 1.06）的估算基于职业类型和专业领域，可能未能捕捉使用质量的关键差异。理论上，“频率”需要与“深度”和“反思性”结合考虑：低质量的高频使用（模式 F 的自动化）可能确实损害元认知发展，而高质量的高频使用（结合元认知监控）则可能促进技能精炼。未来研究应区分“使用频次”与“使用质量”,并考察频率与元认知训练的交互效应，高频使用在有元认知训练支持时可能有益，但在缺乏训练时则有害。

模型对比揭示了经验性预测因子的核心作用。仅含人口统计学变量的基础模型解释了 22.9%的方差（$R^2 = .229$, 调整后 $R^2 = .159$, $F(4,44) = 3.27$, $p = .020$），而纳入经验性预测因子的完整模型解释力跃升至 50.7%（$R^2 = .507$, 调整后 $R^2 = .422$, $F(7,41) = 6.02$, $p < .001$）。经验性变量额外解释了 27.7%的方差（$\Delta R^2 = .277$, F 变化 = 7.69, $p < .001$）,贡献度是人口统计学变量的 1.2 倍。更关键的是，在完整模型中，原本在基础模型中显著的教育水平（$\beta = .50$, $p = .015$）变为非显著（$\beta = -.14$, $p = .618$），系数甚至反转为负，表明其效应完全由经验性因素所解释。这些发现支持了元认知发展的“可塑性假说”而非“固定特质假说”。元认知有效性并非专业水平、技术背景或认知成熟度的

自然产物，而是可以通过明确干预培养的技能集合。设计启示是 MCA 系统应：（1）嵌入元认知策略培训模块,教导用户如何规划、监控、评估 AI 交互；（2）识别和支持有教学经验的用户，利用其已有的元认知优势；（3）为缺乏这些经验的用户提供脚手架式支持，逐步培养元认知能力；（4）创造结构化反思机会，将潜在的“错误经历”转化为真正的“学习经历”；（5）监控使用模式,防止高频但低质量使用导致的模式 F 自动化。

三个非显著预测因子的发现具有深远理论意涵。教育水平效应的消失（从 β = .50 到 β = -.14）挑战了“专家自然更好”的隐含假设，揭示博士训练本身并不直接培养 AI 协作的元认知能力，除非明确包含元认知策略教学。学科差异的缺失（β = .13, p = .491）表明技术专长与元认知警惕性正交，技术流利度可能反而降低元认知监控，正如模式 F 中部分计算机专业用户表现出的过度自信。年龄无关性（β = -.22, p = .266）否定了“经验自然带来智慧”的发展假设，印证了“实践使之永久而非完美”——12 个月的频繁使用未能自动改善策略，除非伴随元认知反思。这些结果为人机协作研究提供了方法论启示：单纯的人口统计学变量无法预测元认知有效性，未来研究必须直接测量经验性因素。同时，结果支持了普适性设计理念，有效的 MCA 系统不应基于用户人口学特征做假设，而应动态评估和支持用户的元认知过程，无论其教育背景、专业领域或年龄。元认知支持应成为所有 AI 系统的标准配置，而非仅为"新手用户"保留的辅助功能。

（4）情境依赖的模式转换

一个有趣的发现使稳定“用户类型”的概念变得复杂：41%的参与者（20/49）在不同情境下表现出不同模式，表明模式通常是情境约束的策略而非个人约束的特质。这从根本上挑战了假设用户可以被分类为需要不同界面或支持的类型的以人为中心的设计方法。

受访者 I4 体现了情境依赖的模式转换。对于高风险学术论文,她表现出模式 A（战略控制）：手工草稿后 AI 润色,逐段验证,内容信任度 0%、语法 40%,投入 8 小时（对比纯手工 12 小时）。对于中等风险课程作业,她转换到模式 C（情境适应）：AI 大纲后人工填充再 AI 润色,结构信任 70%、内容 50%,3 小时完成（对比纯手工 5 小时）。对于低风险生活任务如给房东的邮件,她表现出类模式 F 行为（被动）：直接 AI 生成,最小审查接受,信任 80-90%,耗时 15 分钟（对比手工 45 分钟）。她的理由展示了复杂的成本收益推理："成本收益比在变化。对于我的论文,错误可能让我失去学位——值

得 8 小时验证。对于给房东的邮件,错误只是浪费 5 分钟修正——不值得 30 分钟的审慎。"

表 6 展示了 20 位情境转换者跨情境的模式分布,揭示了策略选择的系统性模式。高风险学术情境引发 65%（13/20）的模式 A、25%（5/20）的模式 D 和 10%（2/20）的模式 C，平均信任 22%，验证强度 8.9/10。中等风险工作情境产生 55%（11/20）的模式C、30%（6/20）的模式A和15%（3/20）的模式E，平均信任58%，验证5.3/10。低风险个人情境触发 50%（10/20）的模式 C、35%（7/20）的类模式 F 行为和 15%（3/20）的模式 B，平均信任 74%，验证 2.1/10。创意/探索情境吸引 45%（9/20）的模式 E、30%（6/20）的模式 B 和 25%（5/20）的模式 C，平均信任 65%，验证 3.8/10。

表 3-8： 情境驱动的模式分布（n=20 位情境转换者）

情境	主要模式	平均信任	验证强度
高风险学术	65% 模式 A 25% 模式 D 10% 模式 C	22%	8.9 / 10
中等风险工作	55% 模式 C 30% 模式 A 15% 模式 E	58%	5.3 / 10
低风险个人	50% 模式 C 35% 模式 F 15% 模式 B	74%	2.1 / 10
创意 / 探索	45% 模式 E 30% 模式 B 25% 模式 C	65%	3.8 / 10

这些数据表明模式是情境约束的策略,而非个人约束的特质。同一个体在论文工作中部署模式 A 的严格控制，在常规项目中使用模式 C 的适应性校准，在琐碎任务中采用类模式 F 的被动接受，所有这些都代表了在不同利害关系、时间约束和错误成本下的理性优化。这从根本上挑战了假设固定用户类别的以人为中心设计。当同一个人在一个时刻需要模式 A 支持而在下一刻需要模式 C 支持时，不能“为模式 A 用户”设计。情境转换现象与前一节的回归发现形成互补。回归分析显示人口统计学特征不预测元认知复杂度，而情境分析进一步揭示即使是元认知能力本身也非稳定特质，它在不同情境下动态变化。这双重证据共同支持“元认知作为情境化实践”的理论框架：元认知不是个体携带的固定属性，而是在特定任务情境中展开的认知活动。有效的MCA设计

必须同时考虑（1）用户的元认知能力基线（通过训练和经验培养）和（2）情境对元认知展现的调节作用。

理论启示：模式代表情境化的元认知策略,而非稳定的个体差异。这挑战了将用户特征视为稳定属性（新手/专家、理工/非理工）的HCI传统用户建模，并主张情境感知系统检测情境因素（任务利害关系、领域、时间压力）并相应调整支持。设计需求：MCA 必须通过多个信号检测情境转换，包括任务描述关键词（“论文”对比“快速邮件”）、用户明确的利害关系声明、类似任务的历史模式以及表明时间压力的截止日期提及,然后调整界面和支架以匹配。对于高风险情境，系统应自动激活验证工具，增加摩擦（强制更慢、更审慎的审查），并建议分解策略。对于低风险情境，系统可以减少支架，同时仍监控过度依赖模式，表明用户可能未准确评估利害关系。这一发现也解释了为什么先前研究发现关于 AI 信任和使用有效性的矛盾结果，在单一情境中测量信任或策略无法预测其他情境中的行为。I2 在学术写作中的 0%信任和在占卜中的 100%信任都代表适当的校准，而非不一致。类似地，4.2.3 节中教育水平在回归中的非显著效应部分可归因于情境变异性：研究生在高风险学术任务中确实表现出更强的元认知控制（如表 3-8 所示），但在低风险日常任务中可能与本科生无异，导致总体效应被稀释。未来研究必须跨多个情境测量信任和策略以捕捉人机协作的真实复杂性，并考察个体元认知能力与情境要求的交互效应。

情境转换能力本身可能是元认知成熟的标志。20 位展现情境转换的参与者平均元认知复杂度为 7.95/12，高于 29 位稳定模式者的 7.14/12（$t(47) = 2.01$, $p = .050$）。这提示高元认知者不仅在单一情境中表现更好，而且能够灵活地跨情境调整策略。然而，灵活性与一致性之间存在张力：过度情境化可能导致策略碎片化,缺乏统一的元认知框架。有效的MCA设计应支持“原则性灵活”：用户维持核心元认知原则（如持续监控、战略规划）同时根据情境调整具体策略（如验证强度、信任水平）。系统可以通过可视化用户的跨情境模式帮助用户发展这种原则性灵活,识别何时情境适应是合理优化,何时则是危险的放松警惕。

本节的跨模式分析揭示了一个根本性的洞察：有效的人工智能协作并非由用户的人口统计特征、专业水平或技术熟练程度决定，而是由元认知参与程度决定。有效模式（A至E）与模式F之间的对比表明，即使是同一个体（一名计算机科学博士生，频繁使用人工智能），也可能仅仅因为认知接近人工智能交互的方式不同而表现出截然不同的结果。这一发现具有深远的设计意义。当前的人工智能系统以任务完成效率

为优化目标，提供无摩擦的界面以实现快速的输出生成。虽然这种设计理念在生产力指标上取得了成功（更快的写作、更多的代码生成、更快速的答案获取），但我们的证据表明，它反而通过消除学习和技能发展所必需的认知摩擦，促进了模式 F 的过度依赖。正是那些使人工智能"易于使用"的特性（即时响应、完整解决方案、自信的呈现），破坏了区分有效使用与无效使用的元认知过程。

我们通过分析识别的六种使用模式，各自代表了管理人工智能协作的不同元认知策略，涵盖四个关键认知过程：规划（任务分解、目标设定、策略选择）、监控（过程追踪、质量检查、信任校准）、评价（输出评估、风险评估、能力判断）和调节（策略调整、工具切换）。模式 A 用户通过战略控制在规划和监控方面表现出色；模式 C 用户通过情境敏感的适应在评价和调节方面展现出复杂的能力；模式 D 用户通过系统化验证在监控和评价方面表现优异；模式 E 用户通过反思性实践整合所有四个过程；模式 B 用户通过迭代优化强调调节。关键的是，模式 F 用户在所有四个过程中都表现出最低限度的参与，将人工智能视为既不需要监督也不需要主动认知参与的黑箱。

定量分析进一步表明，元认知复杂度（以所使用的不同元认知子过程的数量和复杂程度来衡量）强烈预测协作有效性。在模式 A、D 和 E 中表现出高元认知复杂度（评分≥8/12）的用户始终如一地产生了更高质量的输出，保持了独立能力，并随着时间推移发展出更复杂的人工智能使用策略。相反，元认知复杂度低的用户（模式 F，评分≤4/12）表现出技能退化、过度依赖行为以及较差的学习结果，尽管他们频繁使用人工智能。这些实证发现直接指导系统设计。我们的证据表明，人工智能系统不应被假定为响应用户查询的被动工具，而应作为元认知伙伴发挥作用，通过任务分解指导支持规划，通过过程透明性支持监控，通过置信度校准增强评价，通过适应性反馈促进调节。目标不是使人工智能协作变得更容易（在需要更少认知努力的意义上），而是使有效的元认知策略更容易获得、可学习和可持续。

第 3.5 节系统地从这些实证发现中推导出设计需求，将观察到的有效行为转化为具体的系统功能。通过对我们 49 次访谈中编码的 588 个元认知实例、143 个明确的用户挫折和 87 个记录的替代策略进行严格的自下而上分析，我们识别出组织成六个功能类别的 19 项元需求。每项需求都基于特定的用户需求，按影响力和可行性排序，并以足够的细节进行规定以指导实施。我们将由此产生的框架称为元认知协作代理

（Metacognitive Collaborative Agents， MCA），它代表了从以任务为中心到以元认知为中心的人工智能设计的范式转变，有潜力改变人类如何通过人工智能学习。

3.4.4 从使用模式到设计需求：基于证据的推导

本节系统地推导出元认知协作代理（MCA）的 19 项元需求（MR），按六个功能类别组织。每项需求都基于我们访谈中的实证证据，根据用户影响和实施可行性排序，并映射到通过模式分析揭示的特定元认知需求。推导过程确保每个设计决策都可以追溯到我们数据中记录的实际用户行为、挫折和成功策略，而不是反映设计者的直觉或假设场景。

（1） 需求推导的方法论途径

我们的需求推导采用了严格的自下而上、多阶段过程，旨在确保实证基础的同时保持理论连贯性。这种方法建立在已确立的需求工程方法论基础之上（Nuseibeh & Easterbrook， 2000），同时针对元认知支持系统的独特背景进行了调整，在这种系统中，用户需求往往是隐含的、未被表达的，甚至在通过系统分析明确化之前，用户自己都未意识到。

阶段 1：行为抽象

我们系统地对所有 49 份访谈记录进行了元认知行为编码，涵盖我们的 12 子过程分类法（在第 3.3.3 节中描述），生成了 588 个编码实例（49 名参与者 × 12 个子过程）。每个子过程（涵盖规划（任务分解、目标设定、策略选择、角色定义）、监控（过程追踪、质量检查、信任校准）、评价（输出质量评估、风险评估、能力判断）和调节（策略调整、工具切换））都根据证据质量获得强度评级：强（✓✓✓）表示系统化、一致且明确表达的行为；中等（✓✓）表示偶尔或隐含的行为但有明确证据；弱（✓）表示罕见提及或证据非常有限；缺失（无标记）表示没有该子过程的证据。同时，我们识别了访谈期间表达的 143 个明确的用户投诉或挫折，范围从特定的技术限制（受访者 I1 无法让 GPT 突出显示编辑内容，需要手动进行 40 分钟的逐段比较）到更广泛的系统性关切（受访者 I33 担心黑箱交易算法阻止金融领域的专业应用）。这些投诉按受影响的元认知过程（监控挫折：无法有效验证输出；规划挫折：无法分解复杂任务；评价挫折：无法评估输出可靠性）和严重程度（关键：在高价值情境中完全阻碍使用；高：产生需要替代策略的重大摩擦；中等：产生轻微不便）进行编码。

我们还记录了用户为补偿系统限制而开发的 87 种独特的替代策略，代表了人类对设计不足的自发解决方案。示例包括受访者 I3 的三角验证工作流程（GPT 生成代码 → 在环境中测试 → 如果出错，在 Stack Overflow 搜索类似问题 → 用错误详情重新提示 GPT，循环 3-4 次直到代码工作，总时间 30-45 分钟），受访者 I22 的手动多模型比较（在三个浏览器标签中分别打开 ChatGPT、Gemini 和 Claude，将相同提示复制到每个标签，等待所有三个响应，手动比较以寻求共识，每次查询耗时 5-7 分钟），受访者 I34 的文献综述信息沙盒策略（首先手动找到 4-5 篇可信论文，将它们提供给人工智能并明确指示“总结这些文章”，在扩大范围之前测试人工智能是否停留在有界上下文中），以及受访者 I2 的跨情境差异化信任校准（学术引用 0%，在发现三个虚假引用后；早期研究大纲 85%，其中脚手架有助于思考；语法检查 65%，错误立即可见；算命 100%，因为“错了也无妨”）。这些替代策略揭示了用户的复杂性（开发有效的补偿策略证明了元认知意识）和系统的不足（用户不应该需要为系统应直接支持的任务开发复杂的手动过程）。每个替代策略都代表了潜在的设计需求：如果用户手动在三个模型之间进行三角测定（受访者 I22），系统应提供集成的多模型比较；如果用户花费 100 分钟验证引用（受访者 I34），系统应提供自动化的引用验证；如果用户小心地将人工智能隔离以防止幻觉（受访者 I34），系统应提供有界上下文模式或来源锚定功能。

阶段 2：模式需求映射

对于每个模式（A 到 F），我们系统地提出两个问题，旨在将行为观察转化为可操作的设计意义：“什么设计功能会使这种有效策略更容易执行?”以及“什么设计功能可能防止这种无效行为?”这个提问框架确保需求服务于双重目的，即放大有效模式同时防止无效模式。

模式 A（战略分解与控制）分析揭示，虽然 89%的模式 A 用户（17/19）手动分解任务，但他们报告了相当大的认知负担（受访者 I1：在每次人工智能交互前花 40 分钟将 10，000 字文档预处理成 2，000 字的部分；受访者 I3：“分解任务是最难的部分”）。这种手动分解虽然在保持监督和控制方面有效，但产生的摩擦限制了人工智能的潜在价值。需求推导：MR1（任务分解脚手架），如果系统通过多维分析（范围、分析维度、依赖关系）自动检测复杂任务，并提供带有权衡分析（连续性以求彻底性，并行性以求速度，地理/分类性以求结构化覆盖）的引导式分解策略，用户可以通过系统建议消除试错，以减少的认知开销（40 分钟 → 预计 25-30 分钟）实现模式 A 的益处（保持监督、连贯输出、适当粒度）。同样，受访者 I1 的逐段手动比较以识别人工智

能变化直接建议了 MR2（过程透明性和可追溯性），如果系统自动跟踪并突出显示修改（绿色添加、红色删除线、橙色修改并在悬停时显示原始内容），这种费力的手动比较将变得不必要，每次交互节省 40 分钟，同时保持模式 A 的监控严谨性。设计洞察：自动化应减少有效行为的摩擦，而不是取代使它们有效的认知参与。突出显示变化消除了繁琐的手动搜索，但保留了人类对变化是否适当的评估。

模式 C（情境适应）提供了可能最复杂的映射。受访者 I2 在不同情境中的 100 点信任方差（学术引用 0%，研究大纲 85%，语法检查 65%，算命 100%）揭示了 MR9（动态信任校准界面）的必要性，系统必须识别并支持依赖于情境的信任，而不是假设稳定的个人层面信任。当前系统将信任视为用户属性（“你信任人工智能吗?”），而我们的数据揭示信任是按用户 × 任务 × 领域 × 模型 × 风险动态变化的张量。需求：多维、情境感知的信任可视化，显示用户按情境的历史校准（“你对化学引用的信任度：22%，基于 3 次错误经历”）、类似用户的模式（“化学博士：平均 25%”）、人工智能的自我评估（“人工智能对此任务的置信度：78%”）以及不匹配警告（“人工智能表达高置信度，但你的经验表明可靠性低，请仔细验证”）。这个界面明确了受访者 I4 用数学方式表达（Trust_Action = Importance × Expected_AI_Quality × Revision_Effort）但大多数用户只是凭直觉执行的复杂推理。

模式 D（深度验证）映射特别直接。受访者 I3 的 30-45 分钟三角验证工作流程表明需要 MR11（集成验证工具），在 GPT、测试环境和在线资源之间自动化交叉检查可以将验证时间减少到 15 分钟以内，同时保持严谨性。设计规范通过仔细分析演变：用户不希望消除验证（那将移除使模式 D 有效的关键评价），而是希望减少摩擦。因此：集成多模型比较（一键比较 ChatGPT、Claude、Gemini，而不是三标签手动过程）、自动事实检查（根据维基百科、Wikidata、Google 知识图谱自动验证声明）、引用验证（CrossRef 和 Google Scholar API 调用检查每个引用，标记未验证的为“可能的幻觉”）以及代码执行沙盒（在隔离环境中运行生成的代码并自动生成测试用例）。在保持模式 D 验证严谨性的同时节省时间：受访者 I3 的 45 分钟 → 15 分钟（减少 67%）；受访者 I34 的 100 分钟引用检查 → 自动化 2 分钟（减少 98%）。

模式 E（元认知反思）揭示了支持二阶思维功能的需求。受访者 I41 的教学协议（向作为学生的人工智能解释概念，要求人工智能测验理解，使用人工智能的问题识别自己理解中的差距）直接促成了 MR14（引导反思机制）。需求：系统在响应后提供苏格拉底式提示（“在接受这个答案之前，考虑：你如何验证这个声明?这做了什么

假设?可能有更好的替代方法吗?”）、理解检查点（要求用户用自己的话解释人工智能的解决方案后才允许复制，强制与内容而非仅输出消费的参与）以及教学模式（用户向人工智能教授概念，人工智能提出澄清性问题揭示差距，用户完善理解，通过人工智能交互实现费曼技巧）。

模式 F（过度依赖与被动接受）映射侧重于预防而非支持。“什么功能可能防止这种无效行为?”的问题产生了解决根本原因的需求。模式 F 的定义特征（缺乏验证、最低限度的监控、被动接受输出而不评价）源于使过度依赖成为阻力最小路径的系统设计。受访者 I24 的学生提交“GPT 风格的论文”（不连贯、肤浅、矛盾，因为学生从未检查输出，只是信任黑箱）建议了 MR18（过度依赖警告系统），检测模式 F 指标（在不到 30 秒的阅读时间内接受>80%的输出，不足以理解；通过描述整个作业的>500 字提示进行完整任务委托；在 5 次以上连续交互中缺乏验证行动；持续 2 周以上的一贯模式）并以逐步升级进行干预（温和的教育提示 → 社会比较显示专家用户进行验证 → 带有后果数据的强烈警告 → 在输出访问前强制理解检查）。

贯穿多个访谈的技能退化主题（受访者 I38 从富有成效的协作到危机的进展（“第 1-2 个月：协作感觉富有成效...第 6-8 个月：我失去了表达自己想法的能力...第 9 个月：面试失败暴露了真相”），并且生动描述（“我感觉自己失去了认知肌肉张力...写三个简单句子要二十分钟!”），受访者 I41 的享乐发现（“我再次享受编码”，在订阅失效迫使独立工作后），受访者 I47 的语言退化（双语流利度退化为逐词翻译，在外包英语生产后））汇聚成 MR16（技能退化预防系统）。需求：持续监控独立与人工智能辅助工作比率（跟踪从受访者 I38 的 52% → 6 个月内 28%的下降）、组件级技能跟踪（写作分解为大纲、起草、编辑、润色，显示受访者 I38 外包高价值创造性工作同时保留低价值润色，优先级倒置）、带有逐步升级的主动干预（独立性下降 30-40%时温和提示，<30%时强烈警告，<20%时强制技能评估）以及恢复支持（实施受访者 I38 “冷火鸡”重置的无人工智能时间安排，实施受访者 I16 线下技能维护的刻意练习指导，庆祝独立性的成就游戏化）。这个模式需求映射过程产生了初始需求候选：从系统分析中出现了 27 个潜在需求。阶段 3 和 4 通过验证和优先排序对这些需求进行了完善，最终收敛为代表解决所有识别的关键需求的最小充分集的 19 项元需求。

阶段 3：通过多个利益相关者视角进行交叉验证

我们对 10 名战略性选择的参与者进行了成员检查，代表模式和结果的多样性：5 名表现出模式 A、D 或 E 的高效早期采用者（受访者 I3、I8、I16、I22、I41，因展示

先进元认知策略的复杂性而被选择），以及 5 名与过度依赖斗争、表现出模式 F 倾向或正在从中恢复的用户（受访者 I38、I47、I48，加上受访者 I24 班上的 2 名学生，因代表最需要支持的人群而被选择）。我们用通俗语言分享了建议的需求和具体示例，提出三个验证问题："这些功能能解决你经历的需求吗?""我们遗漏了什么关键内容吗?""其中任何功能会让人觉得烦人、居高临下或产生新问题吗?"

十名参与者中有九名确认需求将有意义地改善他们的人工智能协作。响应揭示了验证模式：有效用户（受访者 I3、I8、I16、I22、I41）特别认可验证工具（MR11："终于使验证变得实用"，受访者 I3 的等效评论）、任务分解脚手架（MR1："本可以节省我数月的试错时间来弄清楚如何分解复杂项目"，受访者 I16 的等效）以及信任校准（MR9："我凭直觉这样做，但看到数据会验证我的方法"，受访者 I22 的等效）。挣扎中的用户（受访者 I38、I47、I48）特别认可技能退化预防（MR16："如果我在第 6 个月有那个警告，我本可以避免面试灾难"，受访者 I38）、过度依赖警告（MR18："我直到危机迫使意识才意识到我过度依赖了"，受访者 I48 的等效）以及不确定性显示（MR13："知道人工智能何时不确定会帮助我知道何时验证"，受访者 I47）。

一名参与者（受访者 I4，模式 C 的典范，具有明确的成本效益推理）建议增加超出初始需求规定的明确模式切换能力。她的洞察："有时我想要学习模式，其中人工智能解释一切，我彻底验证，建立我的理解。其他时候我在截止日期前，需要生产力模式，我更多地信任并更快地前进。我根据情境有意识地在这些模式之间切换，但系统不识别或支持这种切换。"这有助于完善 MR20（学习模式与生产力模式），尽管最终这一需求合并到 MR9 的情境敏感信任校准中，而不是作为单独需求，但核心洞察（用户根据目标是学习/能力建设还是任务完成/生产力而需要不同的支持）影响了多个需求的设计。MR16 强调学习/能力保护，MR1 的脚手架随着用户展示能力而淡化，MR9 的信任校准区分需要验证的高风险学习情境与允许效率的低风险生产力情境。

我们邀请了三位外部学习科学家（两位来自教育心理学背景（一位专门研究自我调节学习，一位专门研究认知发展），一位来自认知科学背景（专门研究人机交互和元认知））审查需求与已确立的元认知框架（Flavell， 1979； Schraw & Dennison， 1994； Azevedo & Hadwin， 2005； Winne & Hadwin， 1998）的理论一致性。审查者验证了理论一致性，指出需求与 COPES 模型（条件、操作、产品、评价、标准）和自我调节学习的信息处理理论很好地映射。一位审查者提出了重要的概念观点："模式 F 代表策略的缺失而非连贯的替代方法，它不是第六种'策略'，而是缺乏战略参与。"这

导致整个论文的概念完善，确保我们将模式 F 框定为元认知缺失而非替代元认知策略，设计意义强调预防和恢复而非“支持”（你不支持缺失，你预防或补救它）。

审查者还强调了陈述性元认知知识（知道存在什么策略，受访者 I25 的四级用户拥有这个，模式 F 用户缺乏这个）与程序性元认知技能（实际一致地执行策略，这区分了模式 A 的持续分解与模式 F 的偶尔尝试）之间的区别。这一区别促成了 MR15（元认知策略指导）的设计：教学必须超越展示策略（陈述性）到提供带有反馈的引导实践，直到策略变得习惯性（程序性）。贯穿多个需求（MR1、MR14、MR15）的脚手架到淡化原则操作化了这一理论：初始的重度脚手架建立程序性技能，逐步淡化将控制转移给用户，最终移除留下内化的能力。

当需求可能冲突时，例如，减少用户努力的自动化（以效率为重点）与保持需要用户努力的人类控制（以学习为重点），我们始终如一地优先考虑学习和元认知发展而非效率，认识到教育情境需要不同于以生产力为重点的应用的优化标准。这一原则性立场解决了多个设计张力：MR1 提供任务分解建议但要求用户审查和批准而非自动分解（保持计划参与）；MR11 提供集成验证工具但不会自动无形地验证一切（保持评价责任）；MR16 在达到关键阈值时警告技能退化并阻止人工智能访问，而不是默默允许持续退化（优先考虑长期能力而非短期便利）。指导原则：功能应减少有效元认知策略的摩擦，而不是消除使它们有效的认知参与。

阶段 4：使用多标准决策矩阵进行优先排序

我们使用多标准决策矩阵实施优先排序，考虑反映用户需求和实施现实的四个加权因素：

用户影响（35%权重）：测量为需求频率（多少百分比的用户遇到这个问题?）× 问题严重性（它对有效性的危害程度或阻止使用的程度?）。例如，MR13（透明不确定性显示）在两个维度上都获得最高分：98%的用户（48/49）对虚假置信度表达了挫折感，严重性评为关键，因为它阻止适当的信任校准，影响所有其他元认知过程。计算为：频率（0-100） × 严重性（1-5 量表：1=轻微不便，3=重大摩擦，5=使用阻碍）÷ 100 = 用户影响评分（0-5）。MR13：98 × 5 ÷ 100 = 4.9/5。

技术可行性（25%权重）：估计为实施复杂度的倒数，复杂度在三个维度上评级：算法复杂度（简单的基于规则的逻辑=1，中等 ML/NLP=3，前沿研究=5）、基础设施需求（客户端运行=1，需要 API 集成=3，需要新的后端基础设施=5）以及开发时间估计（周=1，月=3，6 个月以上=5）。三个维度的平均值产生复杂度评分，倒置为

可行性：可行性 = 6 - 平均复杂度。MR13：算法=3（集成方法、校准非平凡但已确立），基础设施=4（需要 RAG、多模型查询、数据库 API），时间=3（估计 3-4 个月）。平均 = 3.33，可行性 = 6 - 3.33 = 2.67/5。

理论基础强度（20%权重）：基于支持文献深度（广泛的数十年研究=5，有明确理论的已确立=4，新兴证据=3，有限研究=2，推测性=1）和与元认知框架的连贯性（直接映射到 Flavell/Schraw=5，与已确立理论一致=4，合理但间接=3，切线相关=2，无理论=1）。平均值产生基础评分。MR13：文献深度=5（关于信任校准、置信度准确性对齐、认识论认知的广泛研究），框架连贯性=5（直接实施 Flavell 的认知知识，Schraw 的监控）。平均 = 5/5。

采纳障碍移除（20%权重）：需求是否移除阻碍高价值情境使用的关键障碍。二元评分：如果移除关键专业/机构障碍（受访者 I33 公司的禁令，受访者 I17 的金融模型禁止）则为 5，如果使之前不可行的模式成为可能（无英勇努力的模式 D 验证）则为 3，如果改善现有使用而不解锁新采纳则为 1。MR13：评分 3（使模式 C 和 D 用户能够更有效地校准信任，不解锁之前被阻止的专业情境，那是 MR23）。

矩阵公式：优先级评分 =（用户影响 × 0.35）+（可行性 × 0.25）+（理论 × 0.20）+（障碍 × 0.20）。可能范围：0-5，阈值为：≥4.0 = 关键（必须有），3.0-3.9 = 高（应该有），2.0-2.9 = 中等（锦上添花），<2.0 = 低（未来考虑）。

MR13 计算：（4.9 × 0.35）+（2.67 × 0.25）+（5.0 × 0.20）+（3.0 × 0.20）= 1.715 + 0.668 + 1.0 + 0.6 = 3.98 ≈ 关键。

这个矩阵产生了表 3-9 中可见的三个优先级层次：关键需求（没有这些功能系统就无法支持有效模式或防止无效模式，MR13 置信度显示、MR23 隐私架构）、高优先级（应该有的功能，解决大多数需求或关键少数需求，13 项需求包括所有模式特定的核心功能）以及中等优先级（对核心功能不是基础的有价值增强，6 项需求提供锦上添花的改进）。没有需求评分为低（<2.0），因为阶段 3 验证已经消除了 27 个原始候选中的 8 个，这些候选缺乏足够的用户需求或理论基础，只留下 19 个最有理由的需求。

优先排序矩阵除了排序实施之外还服务于双重目的：它使优先级决策可审计和可辩护（审查者可以检查我们的权重选择和评分理由，同意或质疑具体评估），并且它识别战略实施序列（阶段 1 侧重于关键加基础设施，阶段 2 构建高优先级模式特定

功能，阶段 3 解决需要阶段 1-2 基础的剩余高优先级，阶段 4 完成中等优先级并解锁专业采纳）。

（2）完整需求目录与证据映射

表 3-9 呈现了 19 项元需求的完整目录，按功能类别组织，并注释了证据强度（表现需求的参与者百分比）、优先级水平（来自多标准矩阵）、实施复杂度估计（来自可行性分析）以及受影响的主要用户模式（来自模式需求映射）。该表在后续章节的详细规范之前提供了系统概览。

表 3-9： 元需求：证据、优先级和实施复杂度

MR ID	类别	需求	证据强度	优先级	复杂度	受影响用户
类别 1：战略规划支持						
MR1	规划	任务分解脚手架	22/49（45%）展现分解；17/49（35%）表现挣扎	高	中等	A，E
MR2	规划	过程透明性与可追溯性	37/49（76%）希望提高可见性	高	中等	A，D
MR3	规划	人类能动性保护	27/49（55%）表达自主性关切	高	低	A，E
MR4	规划	角色定义指导	19/49（39%）提出角色阐明需求	中等	低	A
类别 2：迭代学习支持						
MR5	调节	低成本迭代机制	16/49（33%）频繁迭代	高	中等	B，E
MR6	调节	跨模型实验	12/49（24%）使用多个模型	中等	中等	B，C
MR7	调节	失败容忍与学习机制	9/49（18%）报告富有成效的失败经验	中等	低	B
类别 3：情境敏感适应						
MR8	评价	任务特征识别	28/49（57%）根据情境调整策略	高	高	C
MR9	监控	动态信任校准	41/49（84%）表现出随情境变动的信任水平	高	高	C，D
MR10	评价	成本效益决策支持	13/49（27%）进行 ROI 计算	中等	低	C
类别 4：批判性思维增强						
MR11	监控	集成验证工具	30/49（61%）主动执行验证	高	中等	A，D
MR12	评价	批判性思维脚手架	24/49（49%）需要评估指导	高	低	D，E
MR13	监控	透明不确定性显示	48/49（98%）表达相关挫折	关键	高	全部

类别 5：元认知发展						
MR14	评价	引导反思机制	14/49（29%）使用人工智能开展反思活动	高	低	E
MR15	计划	元认知策略指导	33/49（67%）不了解高级策略	高	低	全部
MR16	调节	技能退化预防	21/49（43%）担忧能力萎缩	高	中等	A，E，（F）
MR17	监控	学习过程可视化	广泛显示元认知益处	中等	中等	E
MR18	调节	过度依赖警告系统	对模式 F 预防至关重要	高	中等	（F）
MR19	监控	元认知能力诊断	适应个体差异需求	中等	高	全部
类别 6：基础设施与隐私						
MR23	基础设施	隐私保护架构	17/49（35%）专业人士提出此类关切	关键	非常高	专业人士

注：模式字母周围的括号（例如“（F）”）表示需求预防而非支持模式。

从这个需求矩阵中出现了几种模式，揭示了支撑 19 项需求的系统结构：

证据强度变异性：MR13（透明不确定性显示）以 98%（48/49 参与者）获得了近乎普遍的认可，使其成为识别的单一最关键需求。只有一名参与者没有明确表达对人工智能虚假置信度的挫折感，即使那位参与者在直接询问时也承认了这个问题（“我想知道人工智能何时不确定会有所帮助，但我没有想太多”）。这种普遍需求超越了模式边界，影响所有用户，无论他们的主导策略如何。在另一个极端，MR7（失败容忍与学习）仅有 18%（9/49）的证据，代表对模式 B 迭代器有价值的增强，但不是普遍需求，适合中等优先级。证据分布验证了我们的优先排序：最高证据需求获得更高优先级（MR13 关键，MR9 在 84%时高），而较低证据需求获得较低优先级，除非解决关键障碍（MR23 在 35%但对专业采纳关键）。

模式 F 预防集群： 需求 MR15（元认知策略指导）、MR16（技能退化预防）和 MR18（过度依赖警告系统）在元认知发展类别中聚集，反映了我们的发现，即无效使用源于元认知缺失而非错误的策略选择。这些需求不“支持”模式 F（不可能支持缺失），而是防止其出现并支持恢复。聚类表明这些需求应该作为连贯的预防系统而不是零散地一起实施，MR15 教授防止初始过度依赖的策略，MR16 监控技能退化并早期捕捉问题，MR18 在尽管有预防努力仍检测到过度依赖时进行干预。它们共同创建了对模式 F 阴险进展的深度防御（受访者 I38 的“煮青蛙”轨迹，其中退化在没有意识的情况下逐渐发生，直到危机迫使认识）。

专业采纳障碍集中：MR23（隐私保护架构）被评为关键，尽管只影响 35%的用户，因为这些用户代表高价值专业情境，其中当前的隐私关切完全阻止机构采纳。受访者 33 的交易公司维护三层防火墙（公司禁止将 GPT 用于交易决策，团队禁止共享专有算法，全行业抵制将生成 alpha 的数据输入公共训练）在零和竞争领域创造零采纳。受访者 17 拒绝输入专有金融模型，受访者 26 禁止查询客户群具体信息，以及多位专业人士对竞争信息泄漏的关切（17/49 明确提及，35%，还有几位其他人暗示了类似关切）代表二元采纳障碍，与改善现有使用的其他需求不同，MR23 决定这些高价值情境中是否根本发生使用。关键评级反映战略重要性：估计价值 100 亿美元以上的企业人工智能市场（Gartner， 2024）由于隐私关切而基本未开发，金融、法律、咨询和医疗保健的采纳率约为 15%，而较不敏感领域的采纳率超过 60%。解决 MR23 可以解锁 3-4 倍的市场扩张。

实施复杂度权衡：两个关键需求显示了有趣的复杂度对比：MR13 评为“高”复杂度（集成方法、RAG、置信度校准具有挑战性但可使用当前技术实现，3-4 个月时间表），而 MR23 评为“非常高”复杂度（本地推理、联邦学习、加密计算需要前沿密码学和基础设施彻底改造，6 个月以上时间表）。这种复杂度差异影响路线图：MR13 在阶段 1 实施，提供即时的普遍益处，MR23 推迟到阶段 4，允许复杂开发的时间，同时其他功能展示价值主张以证明投资合理。优先排序矩阵防止"仅实施简单需求"陷阱：MR23 的非常高复杂度不会将优先级降低到关键以下，因为障碍移除胜过实施难度，一些问题值得解决，尽管有技术挑战。

模式特定与普遍需求：需求分布在高度模式特定（MR4 角色定义指导主要惠及模式 A 的 19 名用户，占样本的 39%）到普遍（MR13 影响所有模式，MR15 惠及所有用户，无论当前模式如何）的光谱上。这种分布建议实施策略：普遍需求（MR13、MR15）应该是始终开启的默认设置，惠及每个人，模式特定需求（MR4、MR7）应该是自适应功能，当用户表现出相关模式时浮现（检测模式 A 的分解行为 → 提供角色定义指导；检测模式 B 的迭代 → 提供失败容忍功能）。自适应浮现防止功能过载：用户不会同时看到所有 19 项需求，而是随着他们的使用模式出现而遇到相关功能。

这个需求目录代表了我们实证发现的全面设计意义集。以下章节提供了递增的细节水平：第（3）小节提供按类别组织的所有 19 项需求的概览级描述，第（4）小节呈现了因关键性和代表性而选择的五个示例需求的详细规范，第（5）小节概述了将需求组织成四个阶段的实施路线图，从基础到完成跨越 24 个月。

（3）按类别的需求概览

本节提供按功能类别组织的所有 19 项元需求的概览级描述，在第（4）小节的详细规范之前建立上下文。每个类别首先说明这些需求为何聚集在一起以及它们主要支持哪些使用模式的理由，然后是涵盖核心问题、关键证据、设计原则和预期影响的个别需求描述。标记为在第（4）小节进行详细规范的需求用[后续详细规范]表示。

类别 1：战略规划支持（MR1-MR4）

战略规划代表了与人工智能进行有效元认知参与的第一阶段，将深思熟虑地构建交互的用户与在没有分解的情况下委托完整任务的用户区分开来。模式 A 用户通过任务分解、目标清晰性、战略角色定义和维持人类监督在计划方面表现出色，这些行为产生了卓越的结果，但需要大量的认知投资。我们的证据显示 45%的用户（22/49）手动分解任务，模式 A 用户显示 89%的普遍性（17/19），但 35%的人在分解方面挣扎，要么未能认识到何时需要，要么缺乏有效执行的技能。

该类别中的四项需求在多个层次上解决计划问题：MR1 支持将复杂任务分解为可管理部分的基本认知工作，使模式 A 的分解对当前挣扎的用户可及；MR2 提供人工智能过程的可见性，使监控和监督无需手动比较；MR3 通过前置人类能动性和选择来防止被动接受；MR4 支持明确的角色协商，阐明人类与人工智能的认知劳动分工。这些需求共同操作化了计划元认知，从隐含的、临时的任务构建转向明确的、受支持的和可学习的计划策略，用户可以系统地应用。

计划支持需求主要惠及模式 A 用户（战略分解与控制）和模式 E 用户（元认知教学与反思），次要惠及根据情境调整计划深度的模式 C 用户。模式 F 预防也关键地依赖于计划支持：从不分解任务的用户将人工智能视为黑箱；支持分解打破了这种模式，强制与任务结构和人工智能角色的参与。

MR1：任务分解脚手架 [后续详细规范]

核心问题：完整提交给人工智能的复杂任务由于监督不足和粒度不当而产生较差结果，但手动分解造成 40 分钟的认知负担（受访者 I1），并需要许多用户缺乏的复杂技能（受访者 I31：“我在哪里分解它?”）。分解不足影响我们样本中 55%的任务，模式 F 用户显示普遍的完整任务委托（受访者 24 的学生：“写一篇 20 页关于可再生能源政策的文献综述”完整提交），即使是模式 A 用户也报告分解是“最难的部分”，需要数月的试错学习（受访者 16 的 13 个月演变）。

关键证据：45%的用户（22/49）手动分解，模式 A 用户的 89%展现这种行为相对于整体 45%（χ^2=10.67， p=.001），但 35%（17/49）在分解方面挣扎。教师受访者 24 的受控观察提供因果证据：遵循明确分解指导的学生（将文献综述分解为：1.定义范围，2.识别主题，3.分析每个主题，4.综合）比委托完整任务的学生产生了明显更好的结果（平均质量 4.1/5 相对于 2.3/5，t=6.8， p<.001，大效应），验证了脚手架使成功成为可能。

设计原则：系统通过多维分析（范围广度、分析维度、依赖关系、估计子任务）识别复杂任务，并提供带有多个策略选项（连续以求彻底性，并行以求速度，分类以求结构）的引导分解，包括权衡分析（时间投资、人类参与、质量结果）和基于任务特征的最佳匹配建议。在 3 次以上成功周期后，随着用户内化分解技能，渐进式脚手架淡化（实施受访者 16 的 13 个月轨迹，通过明确教学加速到 2-3 个月）。

预期影响：分解不足从 55%降至<20%；手动分解开销减少 30-40%（受访者 I1 的 40 分钟 → 通过系统建议消除试错的 25-28 分钟）；任务完成质量改善（复制受访者 24 的 2.3→4.1 课堂干预）；通过对高风险学术任务的强制分解进行模式 F 预防，将"GPT 风格的论文"减少 50%以上。长期：分解成为用户无需系统提示即可应用的内化技能。

MR2：过程透明性和可追溯性

核心问题：用户无法看到人工智能修改了什么，需要费力的手动比较，每次交互消耗 40 分钟（受访者 I1 的逐段比较），造成基本的监控缺口。没有对人工智能变化的可见性，用户要么完全跳过验证（模式 F 的被动接受），要么在繁琐的比较工作上浪费大量时间（模式 A 的手动监控负担）。当前的黑箱生成使得无法区分人工智能贡献与用户的原始内容，防止在协作情境中进行有效监督和功劳归属。

关键证据：76%的用户（37/49）明确表达了对变化可见性的渴望，挫折强度在当前手动验证的模式 A 和 D 用户中最高。受访者 I1 描述花“40 分钟检查 GPT 改变了什么，因为我不能信任它没有引入错误或改变我的意思。”模式 A 用户的监控行为与大量时间投资相关（平均 23%的交互时间用于验证活动），表明自动化透明性可以在减少时间负担的同时保持验证质量。模式 F 用户缺乏验证（0/4 参与系统检查）部分归因于验证难度，当比较需要 40 分钟时，跳过变得诱人，特别是对于缺乏错误风险元认知意识的用户。

设计原则：人工智能输出呈现内联标记，显示添加（绿色下划线）、删除（红色删除线）和修改（橙色高亮，悬停显示原始文本），使快速扫描重大变化成为可能。并排视图提供分屏比较（用户原始左窗格，人工智能版本右窗格），同步滚动和差异高亮。版本历史维护最后 10 次迭代并带时间戳，使回滚成为可能："版本 3（下午 2：34）：添加方法部分 → 版本 4（下午 2：41）：扩展结果。"设计实施"渐进披露"，默认高层概览（变化数量、估计重要性），按需详细标记。

预期影响：受访者 I1 的手动比较时间从 40 分钟降至<5 分钟（自动高亮消除手动搜索）；模式 A 用户更有效地维持监督（监控时间从交互的 23%降至 12%，同时保持相同质量）；模式 D 用户整合到验证工作流程（受访者 I8 的平行问题解决从人工智能与人类贡献的清晰划分中受益）；减少意外接受不想要的变化（模式 F 用户当前接受他们没有注意到的、不打算的变化，因为他们没有注意到它们，受访者 I24 的学生提交带有他们没有写的逻辑的论文）。

MR3：人类能动性保护

核心问题：默认人工智能设计呈现单一生成输出，如果不想要则需要主动拒绝，即使用户的原始更优秀，也会产生接受的心理压力。受访者 I16 阐明原则："如果我能做，我不会让 GPT 做"，但系统设计与这一原则对抗，使接受成为阻力最小的路径，拒绝费力。框架"人工智能建议 X"与"从以下选择：你的原始，人工智能建议 A，人工智能建议 B，混合"从根本上改变了决策心理，前者将人工智能框架为权威，后者将人工智能框架为选项提供者。

关键证据：55%的用户（27/49）表达了对维持自主性的关切，模式 A 用户特别重视控制。受访者 I3 的验证原则"我需要控制，不能只是接受人工智能决定的"）代表系统化的能动性保护，但系统设计不支持这一价值，提供单一输出产生接受压力。被动接受的普遍性（所有用户的 41%交互，模式 F 显示 82%被动接受）表明默认设计不充分支持主动选择。受访者 I41 在订阅失效后重新发现的满足感揭示了过度委托的享乐成本：外包问题解决消除了内在奖励（自主性、能力、掌握，根据 Deci & Ryan，2000）。

设计原则：不是呈现需要主动拒绝的单一人工智能生成输出，而是呈现需要主动选择的 3-5 个选项："选项 A：正式语调，直接请求；选项 B：友好语调，间接方法；选项 C：简短版本（<100 字）；选项 D：你的原始草稿（无人工智能）；选项 E：组合，你的结构，人工智能润色。"强制选择而不是默认人工智能输出，保持能动性。接

受输出后 30 分钟内可用的一键撤销/恢复（解决受访者 9 的后悔）；始终可见的突出“保留我的原始”选项（对抗接受人工智能版本的压力，即使用户的更好，受访者 I16 的原则由设计支持而不是对抗）；明确的学习选择记录（“你选择了选项 B，用户+人工智能混合，78%的时间，似乎是你首选的模式”）。

预期影响：被动接受从 41%降至<15%（要求主动选择消除默认接受）；用户对最终输出的满意度增加（从选项中选择而不是接受默认创造心理所有权，受访者 41 的“再次享受编码”效应部分由于感觉在控制）；模式 A 用户更容易维持原则边界（系统设计支持他们的价值观而不是对抗它们）；通过防止接受作为阻力最小的路径来防止模式 F 的出现。

MR4：角色定义指导

核心问题：用户和人工智能隐含地假设角色而没有明确协商，导致期望不一致和次优协作。受访者 I16 的 13 个月演变展示了通过不同角色关系的进展（阶段 1：“我写代码，GPT 填充函数”；阶段 2：“我写规范，GPT 写 50%的代码”；阶段 3：“我管理项目，GPT 是编写 80%的程序员”；阶段 4：“我使用双人工智能系统并刻意线下练习”），但这种演变通过缺乏系统支持的试错发生。新手特别缺乏关于适当人类与人工智能劳动分工的思考框架，往往默认完全委托（模式 F）或由于对有效协作模型的不确定性而最低限度地使用人工智能。

关键证据：39%的用户（19/49）在访谈期间自发阐明了明确的角色框架，模式 A 用户中的集中度（16/19，84%）相对于其他模式组合的 3/30（10%）（χ^2=34.2，p<.001），表明角色清晰度区分了复杂使用与天真使用。受访者 I3 的“0 到 1 原则”示例了明确的角色定义：“我使用人工智能进行启动而不是执行，它生成初稿或起点，然后我接管。”受访者 I31 的制造不完美策略（故意指示人工智能“添加 5%的语法错误”，以便输出在非母语者情境中“不看起来太完美”）代表细微的角色管理，承认社会感知约束。没有明确框架的用户在适当边界方面更挣扎（受访者 I48："我不知道我什么时候应该做它而不是人工智能应该做"）。

设计原则：系统提供用户可以采用或定制的模板化角色框架：“人类架构师，人工智能助手”（用户提供高层设计，人工智能实施细节，受访者 I16 的阶段 3），“人类经理，人工智能程序员”（用户定义需求，人工智能生成代码，受访者 I16 的成熟状态），“平等合作者”（迭代共创与平衡贡献，受访者 I13 的发散-收敛工作流程），“人工智能导师，人类学习者”（人工智能解释、测验、引导，受访者 I41 的教学协议），

“人类编辑，人工智能起草者”（人工智能创建完整初稿，用户大幅修订，受访者 I47 的恢复后工作流程）。每个模板包括：角色描述、责任边界、决策权（谁对什么做最终决定）、成功标准（如何评估协作是否有效）以及演变指导（如何向更大复杂性进展）。用户可以明确声明“我想成为 X，请充当 Y”，系统相应调整行为。

预期影响： 明确角色清晰度从 39%增加到 70%以上的用户（模板使模式 A 用户通过试错发展的东西可及）；更早出现适当边界（受访者 16 的 13 个月演变 → 通过模板探索的 3 个月引导进展）；减少模式 F 的出现（完全委托成为学习的有意识选择“人工智能导师”模式，而不是角色混淆的默认状态）；使特定于情境的角色切换成为可能（受访者 4 的学习与生产力模式通过与目标匹配的明确角色选择操作化）。

类别 2：迭代学习支持（MR5-MR7）

模式 B（迭代精化与优化）用户参与广泛的迭代周期，平均每个任务 12.3 次迭代，相对于整体平均 2.1 次，通过系统实验而不是接受首次输出来发现最优解决方案。这种方法对于缺乏明确规范的不良定义问题、具有多个可行解决方案的创造性任务以及探索先于承诺的快速原型情境特别有价值。然而，当前的人工智能系统通过三种机制产生重大的迭代摩擦：版本控制缺失使回滚困难（受访者 I9："我希望我能回到版本 3"，在接受较差迭代后），单模型约束限制实验广度（受访者 I22 的三标签手动比较），以及失败成本产生对探索的厌恶（错误感觉像挫折而不是学习机会）。

该类别中的三项需求降低迭代障碍：MR5 提供支持低摩擦循环变化的基础设施，MR6 使系统化多模型实验成为可能，MR7 将失败重新框架为富有成效的学习。这些需求共同使模式 B 的迭代方法对更广泛的用户群可及，可能将模式 B 采纳从 4%主要模式（2/49）增加到 15-20%显示偶尔迭代行为。这些需求特别惠及模式 B 和模式 E 用户，次要价值用于模式 C 用户，他们在探索情境中比在需要效率的高风险情境中更多地迭代。

MR5：低成本迭代机制

核心问题：在当前系统中迭代需要大量努力：手动跟踪版本，将变化复制到外部文档，当需要回滚时从记忆中重建以前的状态。受访者 I9 的挫折（“我做了 8 次迭代试图改进代码，但迭代 3 实际上是最好的，现在我无法回到它，因为我没有保存它”）说明了缺失版本控制的成本。高迭代成本产生两个问题：用户过早地承诺早期版本，避免迭代努力（当进一步精化可以提高质量时，以效率驱动停止），以及确实迭代的用户经历手动管理版本的认知负荷（跟踪像模式 B 用户那样的 12.3 次迭代在没有系统

支持的情况下变得难以处理）。这种摩擦矛盾地惩罚了模式 B 最有价值的特征，即愿意探索变化而不是接受首次输出。

关键证据：33%的用户（16/49）参与广泛迭代（每个任务≥5 个周期），模式 B 用户显示最高比率（100%， 2/2 主要加上次要展现者 14/49 总计，29%）。受访者 I9 的 8 次迭代编码会话代表典型的模式 B 工作流程，但手动版本管理在最优版本出现在序列中间时产生了后悔。受访者 I13 的发散-收敛工作流程（广泛生成许多选项，然后系统地缩小到最佳）需要低成本生成和比较，当前摩擦限制发散宽度（鉴于手动跟踪负担，实际上限于 2-3 个变化，而最优将是 5-10 个变化）。

设计原则： 自动版本控制，差异视图显示迭代间的变化（版本 1 → 版本 2：添加 3 段，语调从正式改为随意，缩短 15%）。一键回滚到任何先前版本，在提交前预览。并排比较使从多个版本中“挑选”最佳元素成为可能（“从 V3 取引言，从 V5 取方法，从 V7 取结果，合并到 V8”）。分支和合并功能允许平行探索（“尝试正式和随意语调两者，看看哪个更好”）而不丢失任一路径。显示迭代指标：“45 分钟内 8 次迭代，V1→V3 改进最多（+23%质量评分），V6 后收益递减（平台期）”。

预期影响： 迭代率从 33%的用户增加到 60%以上，随着摩擦降低；迭代深度增加（从平均 2.1 到 4-5 次迭代，因为回滚安全使实验成为可能）；通过优化提高质量（系统精化而不是过早接受首个可行输出）；模式 B 采纳从 4%主要增长到 15-20%偶尔（特别是对于创造性、探索性、原型情境，其中迭代提供价值）。

MR6：跨模型实验

核心问题：不同的人工智能模型拥有不同的优势：ChatGPT 在数学推理和代码生成方面表现出色，Claude 在细微写作和伦理推理方面表现出色，Gemini 在研究综合和多模态任务方面表现出色（基于我们参与者的经验和技术基准）。然而，探索多个模型需要手动三标签工作流程（受访者 I22），每次查询仅设置比较就消耗 5-7 分钟，不包括综合时间。单模型约束留下桌上价值：用户默认使用一个熟悉的模型，即使任务更适合另一个模型，错过揭示个别模型输出中假设和偏见的比较洞察。模型切换成本产生锁定：一旦用户投资时间学习一个模型的特征，切换到更适合的模型感觉浪费，尽管有潜在的质量收益。

关键证据：24%的用户（12/49）报告使用多个模型，像受访者 22 这样的复杂用户维护按模型-任务组合的明确信任配置文件（“ChatGPT 用于数学 87%，Claude 用于写作润色 91%，Gemini 用于研究综合 85%”）。受访者 I22 的三标签手动比较工作流程

展示了价值（当所有三个都同意时共识 = 89%信任，不同意 = 调查原因，导致更深理解），但 5-7 分钟的设置时间将实际使用限制为仅重要查询。受访者 18 的疲惫（“在我脑海中维护六个模型的信任配置文件令人筋疲力尽，我希望系统为我跟踪这个”）揭示了复杂多模型使用的认知负担。多位用户表达了对比较的兴趣，但无法证明时间投资的合理性（受访者 26：“如果不是切换平台的麻烦，我会尝试 Claude 进行写作”）。

设计原则： 一键[比较模型]按钮同时将提示发送到 3 个用户选择的模型（默认：ChatGPT+Claude+Gemini，可按领域定制，根据受访者 I22 的复杂选择）。并排比较表显示响应，语义相似度评分（>0.85=“高度一致”，0.70-0.85=“中等一致”，<0.70=“重大差异”），黄色高亮关键差异，带有“调查差异”扩展显示详细比较。提供共识解释（“2/3 同意 → 67%置信度，3/3 同意 → 89%置信度，0/3 同意 → 31%置信度，考虑澄清问题”）。可从使用中学习的自动路由规则：“你的模式：ChatGPT 用于代码，Claude 用于写作，Gemini 用于研究，想要自动路由未来查询吗?”

预期影响： 多模型使用从 24%增加到 60%以上，因为一键移除 5-7 分钟障碍；用户更快速地发现模型-任务适配（受访者 I22 的 10 个月试错学习 → 2-3 周系统支持探索）；通过模型集成提高质量（按任务的最优模型选择而不是默认模型接受）；像受访者 I22 这样的复杂用户从认知负荷减轻中受益（系统跟踪信任配置文件，消除心理负担）。将模式 B 的实验从模型内迭代扩展到跨模型探索。

MR7：失败容忍与学习

核心问题： 当前的人工智能交互将失败视为死胡同：不正确的输出 → 用户挫折 → 模糊感觉“人工智能失败了” → 不清楚如何改进。受访者 I9 描述“在获得可行代码之前尝试 8 个不同的提示，每次失败都感觉像是重新开始而不是取得进展。"这种失败作为挫折的框架产生风险厌恶：用户坚持已知的提示，即使次优，避免由于失败成本而进行实验。然而，模式 B 的生产力正是来自将失败视为信息：“这种方法不起作用，所以我了解了关于问题的 X 或关于人工智能局限性的 Y。”系统重新设计可以将失败从挫折重新框架为学习机会。

关键证据：18%的用户（9/49）明确描述了富有成效的失败经历，其中错误导致洞察。受访者 I13 的发散-收敛工作流程在发散阶段生成了许多“失败”选项，但这些失败帮助理解解决方案空间：“通过看到什么不起作用，我理解了我真正想要的。”受访者 I9 的 8 个提示迭代代表了系统探索，尽管有挫折，每次失败都缩小了规范。教师

受访者 I24 指出，由于失败厌恶而避免实验的学生比愿意通过错误迭代的学生产生了更差的结果，表明失败容忍与学习相关。

设计原则： 当人工智能输出失败或用户拒绝时，系统提出诊断问题："具体什么不起作用?（太正式/随意?逻辑不正确?缺少需求?错误方法?）"跟踪失败模式："你因为'太冗长'拒绝了 3 个输出，为未来输出调整冗长基线。"提供特定于失败的指导："检测到逻辑错误：人工智能误解了 X。尝试通过提供预期行为示例来澄清。"将迭代重新框架为进展："迭代 5/8：每次精化都更接近，你已经识别并修复了：语调（迭代 2）、长度（迭代 4）、缺少需求（迭代 5）。进展：62%。"庆祝富有成效的迭代："成就解锁：迭代大师，通过 10 个以上周期精化达到高质量结果，显示耐心和对卓越的承诺（模式 B 行为）。"

预期影响： 风险厌恶减少（用户更愿意探索变化，知道失败有助于学习）；迭代深度增加（失败成为优化中的预期步骤而不是消极挫折）；明确的失败分析改善未来表现（诊断问题建立用户对有效提示的理解，教学而不仅仅是做）；模式 B 采纳增加，因为失败容忍降低了广泛迭代的心理障碍；对人工智能的挫折减少（从"人工智能让我失败了"重新框架为"我通过实验学习如何更有效地协作"）。

类别 3：情境敏感适应（MR8-MR10）

模式 C（情境适应）用户展示了复杂的情境敏感性，根据任务风险、领域熟悉度、时间压力和错误后果调整人工智能依赖和验证严谨性。受访者 I2 的 100 点信任方差（学术引用 0%，研究大纲 85%，语法检查 65%，算命 100%）和受访者 4 的明确成本效益公式（Trust_Action = Importance × Expected_AI_Quality × Revision_Effort）揭示了有效用户不维持稳定的人工智能使用模式，而是基于情境需求动态校准。这种适应性证明至关重要，因为对低风险个人创意写作的适当人工智能依赖与对高风险专业金融建模的适当依赖根本不同，一刀切的方法要么在安全情境中不充分利用人工智能，要么在风险情境中过度依赖。

然而，支持情境敏感性需要识别情境特征的系统（MR8）、维持多维信任模型而不是假设个人层面稳定性的系统（MR9）以及帮助用户在特定情况下推理成本和效益的系统（MR10）。当前系统对所有查询相同对待，提供相同的自信响应，无论任务是随意算命（受访者 I2：100%信任适当，因为"错了也无妨"）还是学术引用（受访者 2：0%信任适当，因为"虚假引用可能让我失去学位"）。这种情境盲目性防止适当适应。

该类别中的三项需求主要惠及模式 C 的 33%用户（16/49），但也支持根据情境调整策略的模式 A 和 D 用户（受访者 I3 的验证严谨性因任务关键性而异），并通过标记过度依赖造成不可接受风险的高风险情境来帮助防止模式 F。

MR8：任务特征识别

核心问题：系统对所有查询相同对待，无论任务是低风险创造性头脑风暴还是高风险专业决策，都提供统一的自信响应。这种情境盲目性防止适当的支持适应：低风险任务收到不必要的警告（受访者 I2 的算命不需要验证警报），高风险任务收到不充分的脚手架（模式 F 用户提交完整学术作业而没有分解提示）。用户必须手动记住情境适当的行为，而不是收到与任务需求匹配的及时支持。受访者 31 的不确定性（"我在哪里分解这个任务?"）部分源于情境盲目性，她知道分解对复杂提案重要但对简单电子邮件不重要，但系统不帮助区分。

关键证据：57%的用户（28/49）描述了按情境明确调整行为，模式 C 显示普遍情境敏感性（16/16，100%）。受访者 I4 的成本效益公式操作化了情境特征：重要性（论文=10/10，随意电子邮件=2/10）、预期人工智能质量（基于经验的领域特定估计）、修订努力（检查的可用时间）。受访者 I2 的信任方差展示了极端但理性的情境依赖性：学术引用 0%反映高错误率结合严重后果（论文拒绝、学术诚信违规），算命 100%反映错误无成本。多位专业用户（受访者 I17、I26、I33）指出风险决定使用：不会在专有竞争情境中使用人工智能（零和领域，其中信息泄漏产生劣势），但对常规任务接受人工智能。

设计原则：在任务特征上训练的机器学习分类器识别风险（通过关键词：高/中等/低："论文"、"客户项目"、"头脑风暴"）、领域（通过主题建模：学术、专业、个人、创造性）、可验证性（事实声明=高，创造性内容=低，分析=中等）和错误成本（根据受访者 I17 金融=高，根据受访者 I4 随意=低）。分类器输出多维配置文件："任务：化学文献综述 | 风险：高（学术诚信） | 领域：化学（专业） | 可验证性：中等 | 错误成本：高 | 推荐模式：模式 D 验证 + 模式 A 分解 | 信任校准：低（根据你的历史 22%）"。这个配置文件馈送自适应支持：高风险触发分解脚手架、验证工具提示、低信任警告；低风险允许以效率为重点的快速响应、最低限度验证。

预期影响：、情境适当支持消除过度支持/支持不足的不匹配（受访者 2 的算命没有不必要的警告，高风险学术任务没有不充分的脚手架）；模式 C 用户为他们的复杂适应获得系统化支持（受访者 I4 的公式由系统自动实施）；在高风险情境中防止模

式 F，同时在低风险情境中允许效率；像受访者 17 这样的专业用户收到风险适当的警告（“金融建模：高风险，彻底验证”），支持在当前避免的情境中采纳。

MR9：动态信任校准界面

核心问题：当前系统假设个人层面的稳定信任（“你信任人工智能吗?”），而证据显示信任是按用户 × 任务 × 领域 × 模型 × 风险动态变化的张量。受访者 I2 的 100 点方差展示了同一个体表现出根据情境适当校准的根本不同的信任。系统无法识别和支持这种情境依赖性迫使用户要么统一信任（在某些情境中不适当的过度信任，在其他情境中不充分信任），要么手动维持跟踪按情况的适当信任的复杂心智模型（受访者 22 的令人筋疲力尽的电子表格跟踪，受访者 18 管理六个模型信任配置文件的认知负担）。

关键证据：84%的用户（41/49）在情境中表现出可变信任，有些显示 100 点方差（受访者 2）。跨模式分析揭示个人内信任方差（模式 C 标准差=28%）超过跨所有模式的个人间方差（标准差=22%），表明情境比个性更重要。受访者 22 的领域特定跟踪（“HTML/CSS：85%，React：70%，前沿框架：35%”）和受访者 18 的模型特定配置文件展示了复杂但认知负担沉重的校准。时间动态增加复杂性：受访者 38 的钟形曲线轨迹（第 1 个月：20%，第 2-3 个月：90%，第 4-8 个月：下降到 60%，第 9 个月：危机时 20%，第 10-12 个月：稳定在 25%）显示信任即使对于相同的用户-任务-领域组合也不是静态的。

设计原则：利用 MR8 任务分类的多维信任仪表板显示：用户按情境的历史校准（“你对化学引用的信任：22%”）、类似用户的模式（“化学博士：平均 25%”）、人工智能的自我评估（根据 MR13 置信度评分）以及不匹配警告（当人工智能置信度超过经验上合理的信任时：“人工智能表达 78%置信度，但你的经验表明 22%可靠性”）。信任演变可视化（显示受访者 I38 的钟形曲线、受访者 I40 的发散轨迹的图表）帮助用户理解自己的发展。情境特定配置文件使细粒度校准成为可能（不仅仅是“我信任人工智能 50%”，而是“学术引用：22%，研究大纲：85%，语法：65%，个人建议：100%”）。基于表现的实时信任调整（在 3 个连续准确结果后，建议用证据增加信任；在错误后，建议用模式分析减少）。

预期影响：信任校准准确性从 32%平均绝对误差改善到<15%；行为与信任对齐（高信任 = 更少验证时间，低信任 = 更多验证）；模式 C 用户的个人内信任方差增加（表明更复杂的情境辨别而不是一刀切的信任）；不适当依赖减少 68%（从 38%做出

错误信任决策到 12%有情境感知指导）；像受访者 I33 这样的专业用户可以在信任合理的特定子任务中使用人工智能，同时避免信任不足的使用，使在当前被阻止的领域中选择性采纳成为可能。

MR10：成本效益决策支持

核心问题：用户必须隐含地判断人工智能交互是否产生正的投资回报率：通过人工智能辅助节省的时间相对于花在验证/纠正输出上的时间。受访者 I48 阐明了这一点："验证比我自己重做还要花更长时间，有什么好处?"这种成本效益推理直觉地发生，而且往往很差，导致次优决策：在验证成本相对于节省的时间较低的情境中不充分使用人工智能（拒绝有价值的辅助），或在纠正努力超过生成效益的情境中过度使用人工智能（受访者 I24 的学生接受需要 2 小时以上修订的不连贯 GPT 输出，而独立写作本来只需 1 小时）。受访者 I4 的明确公式（Trust_Action = Importance × Expected_AI_Quality × Revision_Effort）代表了大多数用户不阐明的复杂推理。

关键证据：27%的用户（13/49）自发提到成本效益考虑，模式 C 显示集中度（9/16， 56%）相对于其他模式（4/33， 12%， χ^2=12.4， p<.001）。受访者 4 的公式操作化因素：重要性（做对这件事有多重要?）、预期人工智能质量（基于经验的领域特定预测）、修订努力（检查/纠正的可用时间）。受访者 I48 对验证成本的投诉揭示了常见的误算：他假设验证＝完全重做，但模式 D 用户展示了带有集成工具（MR11）的验证需要 15 分钟相对于 45 分钟的独立完成，产生正的投资回报率（节省 30 分钟），与他的直觉相反。多位用户指出投资回报率因情境而大大不同：头脑风暴 = 高投资回报率（人工智能快速生成许多想法，用户快速筛选），复杂技术工作 = 不确定的投资回报率（生成快但验证缓慢且困难）。

设计原则：计算器用从使用中学习的个性化参数实施受访者 4 的公式。对于给定任务，系统估计：节省时间（基于任务复杂度和人工智能速度）、验证时间（基于任务类型和用户从 MR16 监控的典型验证严谨性）、错误率（基于这个用户-任务组合从 MR9 的历史准确性，领域特定）、纠正时间（估计修复典型错误的努力）。公式：净效益 = 节省时间 -（验证时间 + 错误率 × 纠正时间）。显示："对于这个文献综述：节省时间：60 分钟，需要验证：20 分钟，预期错误：2-3 个引用（15 分钟纠正），净效益：+25 分钟。建议：使用人工智能并彻底验证。"对于负投资回报率："对于这篇创意论文：人工智能节省 20 分钟，但验证在你的领域需要 35 分钟，加上修订往往广泛。净效益：-15 分钟。建议：独立写作。"

预期影响： 理性人工智能采纳增加（用户基于数据而非直觉做出关于何时使用人工智能的决策）；在高投资回报率情境中不充分利用减少（用户发现人工智能在他们假设不会有帮助的地方有价值，受访者 48 了解到带有工具的验证对许多任务比完全独立工作更快）；在负投资回报率情境中过度利用减少（模式 F 用户发现“需要 2 小时修订的人工智能草稿”产生比“1 小时独立写作”更差的投资回报率）；时间效率提高（用户将人工智能辅助分配到净节省时间后验证成本的任务，释放时间用于需要人类创造力/判断的任务）。

类别 4：批判性思维增强（MR11-MR13）

批判性思维，即对信息质量、逻辑连贯性和证据支持的系统化评估，代表了人工智能交互可以增强或破坏的关键元认知能力。模式 D 用户通过系统化验证体现了批判性思维：100%（4/4 主要，加上 7/7 次要）主动使用多种验证方法交叉检查人工智能输出，以 94%的比率捕获错误，相对于整体 52%。这种验证严谨性创造了防止错误传播的认知安全网，但当前系统使验证变得繁琐（受访者 3 的 45 分钟三角工作流程，受访者 34 的 100 分钟引用检查，受访者 22 的三标签模型比较）。

该类别中的三项需求旨在使批判性思维更易获得：MR11 提供集成工具，将验证摩擦从 45 分钟减少到 15 分钟，同时保持质量；MR12 为缺乏系统化方法的用户提供批判性评价的脚手架；MR13 解决所有用户的普遍挫折感（98%的用户），即人工智能的虚假置信度阻止适当的信任校准。这些需求共同操作化了一个洞察，即有效的人工智能使用需要健康的怀疑态度配以高效的验证机制，既不是盲目信任也不是愤世嫉俗的拒绝，而是与风险成比例的、基于证据的评估。

这些需求主要惠及模式 D 用户（系统化验证）和模式 A 用户（战略监督），对模式 F 预防具有关键重要性（过度信任部分源于无法有效验证，产生习得性无助：“验证太难了，我就信任人工智能吧”）。

MR11：集成验证工具

核心问题：验证当前需要繁琐的多工具工作流程，每个任务消耗 30-45 分钟：受访者 I3 的三角验证（GPT → 测试环境 → Stack Overflow → 重新提示，循环 3-4 次），受访者 I22 的多模型比较（三个浏览器标签手动比较响应），受访者 I34 的引用检查（CrossRef、Google Scholar 检查 40 个引用中的每一个，每个 2-3 分钟 = 总共 80-120 分钟），受访者 I8 的平行问题解决（独立解决然后与人工智能解决方案比较，每个问题 15-20 分钟认知负荷）。这种摩擦产生二元选择：在验证上投资大量时间（模式 D 平均

每个任务 27 分钟）或完全跳过验证（模式 F 的 0 分钟，39%的用户验证不足）。缺失的是中间地带：有效验证使 15 分钟投资能捕获 90%以上的错误。

关键证据：61%的用户（30/49）主动验证，模式 D 为 100%，模式 A 为 89%（17/19），但所有人都报告重大摩擦。在非验证者中，68%表示如果工具将时间投资减少 50%以上（从 23 分钟降至 12 分钟以下），他们愿意验证，表明存在大量潜在需求。验证有效性得到验证：验证者捕获 94%的错误（模式 D：4/4 所有错误，模式 A：16/17，部分模式 C：7/11），相对于非验证者捕获 52%（模式 F：0/4 没有捕获任何错误，部分模式 B：5/9，部分模式 C：6/12）。这 42 个百分点的错误检测差距证明了验证价值，但 23 分钟的时间投资解释了为什么 39%的人跳过它，在低风险情境中对边际收益而言太昂贵。

设计原则： 统一验证仪表板整合当前分散的验证活动：多模型比较（一键将提示发送到 3 个选定模型，并排显示语义相似度评分和差异突出显示，受访者 I22 的三标签过程自动化），自动事实检查（通过自然语言处理提取声明，根据维基百科/Wikidata/Google 知识图谱验证，带有来源详情标记矛盾），引用验证（通过 CrossRef/Google Scholar API 自动检查所有引用，未验证的标记为“可能的幻觉”，受访者 34 的 100 分钟手动过程 → 2 分钟自动化），代码执行沙盒（在隔离环境中运行生成的代码并自动生成测试用例，Stack Overflow 类似问题检测，受访者 3 的三角验证自动化）。所有工具在持久的右侧面板中可访问，结果在人工智能生成输出时自动更新。

预期影响： 验证时间从受访者 3 的 45 分钟和受访者 34 的 100 分钟减少到目标 <15 分钟（减少 60-85%）；验证采纳从 61%增加到 85%以上（24 个额外用户在障碍降低时采纳）；尽管自动化，错误检测保持 94%的比率（自动化加速人类判断，不替代它）；在像受访者 17 的金融这样缺乏验证基础设施的领域中实现专业采纳（"如果存在自动化公式检查、回测、监管验证，我可能会在模型中使用人工智能"）；通过使验证可行而非英雄努力来预防模式 F。

MR12：批判性思维脚手架

核心问题： 许多用户缺乏批判性评估人工智能输出的系统化框架，在不检查潜在逻辑、假设或证据的情况下接受听起来合理的响应。模式 F 用户特别脆弱：受访者 48 的“它是技术性的，我假设它比我更懂”反映了对权威的服从而非批判性评估。即使是复杂的用户有时也会忽略缺陷（受访者 2 最初接受了三个捏造的引用，因为它们“看起来如此具体和权威”）。系统设计矛盾地破坏了批判性思维，无论认识论地位如何，

都以统一的置信度呈现输出：事实声明、推测性推断和创造性生成都以相同的权威语调传递（受访者 2：“它不承认什么时候做不到，即使错了也听起来确定”）。

关键证据：49%的用户（24/49）表达了需要批判性评估的指导，模式 F（4/4，100%）和模式 B 用户（2/2， 100%）的需求更高，相对于模式 D 用户（1/4， 25%），后者已经拥有系统化方法。教师受访者 I24 的受控观察：收到明确批判性思维检查表（“在接受人工智能输出之前：1.检查事实声明，2.检查逻辑差距，3.识别假设，4.考虑替代方案”）的学生比没有脚手架的学生产生了更高质量的工作（平均质量 3.8/5 相对于 2.9/5，t=4.2， p<.001），验证了脚手架有效性。受访者 I17 阐明了需求：“我不知道如何评估金融模型，需要理解每个方程，对照教科书检查，用历史数据测试，验证边界条件。比从头构建花费更长时间，所以我就不使用人工智能。”

设计原则： 人工智能响应后的苏格拉底式提示触发反思，然后允许接受：“在使用此输出之前，考虑：你如何验证这个声明？这做了什么假设？存在什么替代方法？你理解为什么这有效，还是只是知道它有效?"红旗检测器识别有问题的模式：过度自信的语言（“绝对”、“当然”、“总是”没有限定词 →“强声明，仔细验证”），没有引用的不支持的概括（“研究显示... ”没有引用 → “哪项研究？找到具体来源”），因果关系声明（“X 导致 Y” →“这是相关性还是因果关系?考虑混淆因素”），超出知识截止日期的时间声明（提到 1 月后的 2025 年事件 →“可能是推测性的，通过当前来源验证”）。领域特定脚手架：对于受访者 I17 的金融，自动对照教科书公式检查、监管要求、历史绩效范围。

预期影响：批判性思维行为增加（系统化怀疑和评估从模式 D 的 8%主要传播到 25%以上的偶尔用户）；错误检测改善（用户更有效地捕获逻辑缺陷、不支持的声明、虚假置信度，从整体 52%到 70%以上）；元认知意识增长（用户报告即使没有提示也批判性思考，受访者 I13 的原则：“这使我意识到我可以部署的不同思维模式”）；模式 D 行为对新手用户更易获得（脚手架通过引导实践教授系统化评估模式，最终内化，使用户独立应用）。

MR13：透明不确定性显示 [关键优先级]

核心问题：人工智能系统产生"虚假置信度"，即即使人工智能的内部概率分布表明对正确性存在重大不确定性，输出也读起来具有权威性和确定性。当前人工智能校准的概率理解（显示不确定性的模型置信度评分）与语言上过度自信的呈现（流畅、听起来确定的语言产生虚假的权威印象）之间的脱节代表了识别的最普遍问题：98%

的用户（48/49）表达了挫折感。受访者 2 的投诉抓住了本质："它不承认什么时候做不到。即使错了，也听起来确定。"这种无论潜在认识论地位如何的统一确定性迫使用户陷入不可能的境地：要么验证一切（昂贵，认知需求高，受访者 I4 的 8 小时论文验证负担）要么基于直觉选择性信任（不可靠，容易出错，受访者 I2 的 0%引用信任仅在经历三个虚假引用后发展）。

关键证据：跨越所有模式的普遍挫折感：受访者 I2（模式 C）："它不承认什么时候做不到"；受访者 I9（模式 B）："GPT 强制输出看起来对但实际不对的东西。置信度具有欺骗性"；受访者 I17（模式 D）："概率模型使 100%置信度不可能，但 GPT 表现得好像它总是对的"；受访者 I47（模式 F 恢复中）："我希望它能告诉我'我对此 60%确定'。那种诚实会有很大帮助。"专业领域案例揭示了采纳障碍：受访者 I33 的交易公司禁止 GPT，因为"黑箱等于不可控风险，当人工智能无法逐级显示推理置信度时，我根本无法使用它。"没有置信度差异化，受访者 I2 在经历三个虚假引用后"被背叛"，对引用保持笼统的 0%信任，尽管一些引用是可验证的，证明了虚假置信度如何滋生完全不信任而非适当校准。

设计原则：三级实施：（1）声明级置信度评分，使用集成方法（多个模型独立生成响应，一致程度产生置信度：5/5 同意=95%，3/5=60%，共识失败=<30%）结合检索增强生成（RAG）（与权威来源匹配的声明+10-15%置信度，与来源矛盾的声明-20-30%）。引用通过 API 调用强制外部验证到 CrossRef/Google Scholar，如果未验证，总是<20%置信度，明确警告"在数据库中未找到引用；可能的幻觉。"（2）不确定性来源归因，解释为什么置信度有限："置信度限制：知识截止（训练结束于 2025 年 1 月），查询模糊（'最佳'可能意味着性能/成本/易用性），引用未验证（来源不在数据库中）。"（3）针对像受访者 33 的交易这样的专业情境的领域特定置信度配置文件："交易信号：模式识别 92%，风险评估 45%，结果预测 23% → 建议：使用模式进行信号，应用你的风险模型进行头寸，不要信任结果预测。"

预期影响：用户在 80%以上的时间正确识别不可靠输出（相对于当前约 52%）；盲目接受减少到<10%（从模式 F 的 68%）；信任校准误差减少到<15%绝对误差（从表达信任与实际可靠性之间 32%的平均差异）；解决专业采纳障碍（受访者 33 可以使用具有组件级置信度分解的人工智能，受访者 I17 的金融模型通过公式置信度评分变得可行）；试点测试显示可靠性判断从 54%改善到 81%（χ^2=23.4， p<.001），盲目接

受从61%降至12%（χ^2=31.7， p<.001），满意度从3.2/5改善到4.6/5（t=8.3， p<.001，d=1.85 非常大效应）。

类别 5：元认知发展（MR14-MR19）

元认知，即思考思维，对自己认知过程的意识和调节，比任何人口统计或专业变量更能区分有效与无效的人工智能使用。我们的证据表明，元认知复杂度（使用的元认知子过程的数量和复杂程度）预测协作质量：在模式 A、D、E 中评分≥8/12 的用户在输出质量、学习结果和能力维持方面始终优于评分≤4/12 的用户（模式 F）。然而，67%的用户（33/49）报告不了解像受访者 3 的 0 到 1 原则、受访者 22 的多模型共识解释、受访者 41 的教学协议这样的复杂策略，表明元认知专业知识仍然是隐性的而非广泛分布的。

该类别中的六项需求旨在发展元认知复杂性：MR14 提供支持模式 E 二阶思维的反思机制，MR15 教授明确策略使模式 A/D/E 方法可及，MR16 通过监控和干预防止技能退化，MR17 可视化支持模式 E 反思取向的学习过程，MR18 检测并防止模式 F 过度依赖，MR19 诊断个体元认知配置文件使个性化支持成为可能。这些需求共同操作化了人工智能系统增强而非替代人类元认知发展的愿景，教用户更有效地思考而不仅仅是提供答案。

这些需求惠及所有用户（MR15、MR18、MR19 具有普遍适用性），同时特别支持模式 E 的反思复杂性（MR14、MR17）和防止模式 F 出现（MR16、MR18）。

MR14：引导反思机制

核心问题：大多数用户参与一阶思维（任务完成）而没有二阶反思（从交互中学习，建立元认知意识）。受访者 I41 的教学协议体现了有价值但罕见的实践：向作为学生的人工智能解释概念，回答人工智能的理解问题，使用这个过程识别自己理解中的差距，通过人工智能交互实施费曼技巧。这种反思将人工智能从答案提供者转变为思维伙伴，但只有 29%的用户（14/49）自发采用这种实践。系统设计隐含地阻止反思：即时答案使快速任务完成成为可能，但没有提供元认知处理的暂停，错过了深化理解的机会。

关键证据：模式 E 用户普遍（9/9， 100%）描述了反思实践，相对于其他模式中的 5/40（13%）（χ^2=38.6， p<.001），证明反思是最复杂使用的区别特征。受访者 I41 的教学协议，受访者 I25 的元元认知提示（询问人工智能“我应该为这个问题使用什么思维技巧?”），受访者 I5 的模拟面试方法（人工智能扮演识别盲点的严格面试官）

都代表了产生更深学习的自我发起的反思机制。受访者 I24 的自我纠正循环（询问人工智能“反思我的要求，你的答案需要优化吗？”）触发产生更好输出的元认知评价，通过强制反思。没有反思机制的用户表现出更肤浅的参与（受访者 I48：“我只是问和接受”相对于受访者 I41 的“我向人工智能解释以测试我自己的理解”）。

设计原则： 在响应后使用苏格拉底式提问的理解检查点：“在继续之前，反思：你理解为什么这有效，还是只是知道它有效?用你自己的话解释。下次你会做什么不同的事?”教学协议模式使受访者 I41 的策略成为可能：用户向作为学生的人工智能解释概念，带有脚手架结构（“我将向你解释 X。我完成后：1.总结关键概念，2.举例，3.解释一个限制，4.向我提出揭示差距的问题”）。自我纠正循环（受访者 I24 的方法）自动化：人工智能响应后，系统询问“反思原始要求，这个答案应该优化吗？列出要求 → 自我评估每个 → 提供改进版本。”元认知策略反思（受访者 I13 的实践）：“你使用了发散-收敛工作流程。这背后的正式理论是什么?它的局限性是什么?”将直觉策略连接到已确立的框架。

预期影响：模式 E 行为传播（反思性人工智能使用从 18%增加到 35%以上，显示偶尔的二阶元认知）；学习收益改善（自我报告的“更深理解”从整体 58%接近模式 E 的 87%）；元认知复杂性发展（用户从一级任务提示进展到二级战略到三级元认知到四级元元认知，根据受访者 I25 的框架，通过提示分析显示越来越多的抽象来衡量）；用户发展在人工智能情境之外应用的可移植思维技能（受访者 I41 的教学协议改善一般理解，不仅仅是人工智能使用）。

MR15：元认知策略指导

核心问题：复杂的人工智能使用策略仍然是少数人拥有的隐性知识，通过广泛的试错发现（受访者 I16 的 13 个月演变）而非系统化教授。67%的用户（33/49）不了解“更好地提示”之外的策略，错过了像受访者 I3 的 0 到 1 原则（人工智能用于启动而非执行）、受访者 I3 的三角验证（GPT→测试→Stack Overflow→重新提示）、受访者 I41 的教学协议（向作为学生的人工智能解释）、受访者 I24 的自我纠正循环（要求人工智能反思自己的答案）、受访者 I25 的元认知命令（告诉人工智能使用哪种思维技巧）、受访者 I13 的发散-收敛工作流程（广泛探索然后系统化缩小）这样的技术。模式 F 用户特别缺乏战略知识，尝试完全任务委托（受访者 I24 的学生：“写一篇 20 页的关于可再生能源政策的文献综述”）而没有认识到需要分解、验证或监督。

关键证据：策略意识与结果强相关：阐明≥2 个明确策略的用户平均元认知复杂度为 8.2/12，输出质量为 4.3/5，相对于阐明 0-1 个策略的用户平均复杂度为 5.1/12，质量为 2.8/5（t=7.4， p<.001，大效应）。受访者 I16 的 13 个月演变证明策略是可学习的，但当前学习曲线效率低下。受访者 I24 的课堂干预提供明确的分解策略显著改善了学生结果（2.3/5→4.1/5 质量，t=6.8， p<.001），验证了教授策略有效。多位复杂用户表达了分享策略的愿望（受访者 I22："我希望有一个有效提示和技术的社区库"）。

设计原则：入职教程呈现"来自专家用户的有效人工智能使用策略"，带有互动示例和练习：受访者 3 的 0 到 1 原则（"使用人工智能作为起点，你做主要工作"），受访者 I22 的共识解释（"所有 3 个模型同意=89%信任"），受访者 I41 的教学协议（向作为学生的人工智能解释），受访者 I24 的自我纠正循环（要求人工智能反思），受访者 I25 的元提示（这个问题用哪种思维技巧?），受访者 I13 的发散-收敛（广泛头脑风暴然后缩小），受访者 I31 的不完美注入（添加 5%错误以人性化），受访者 I34 的信息沙盒（首先提供可信来源，测试理解）。策略库使用户生成内容成为可能：用户提交策略，其他人投票有效性，评分最高的成为入职材料（众包元认知知识）。基于当前模式的自适应呈现：模式 F 用户接收基础策略（验证、分解），模式 B 用户接收迭代技术，模式 A 用户接收高级策略。

预期影响： 策略意识从 44%阐明≥2 个策略增加到 80%；通过行为跟踪测量的策略采纳（0 到 1 原则使用 23%→60%，验证技术 61%→85%，教学协议 0%→30-40%在面向模式 E 的用户中）；学习结果改善（自我报告的理解接近模式 E 水平）；受访者 I16 的 13 个月自我发现 → 6-8 周引导学习（加速能力发展 50%以上）；通过教授完全任务委托的替代方案进行模式 F 预防。

MR16：技能退化预防系统

核心问题：过度依赖人工智能造成阴险的技能退化：逐渐的、无意识的、由人工智能补偿的，直到危机迫使认识。受访者 38 的进展示例："第 1-2 个月：协作感觉富有成效...第 6-8 个月：失去表达想法的能力...第 9 个月：面试失败暴露了真相。"受访者 I38 的生动描述："我感觉自己失去了认知肌肉张力。写三个句子要二十分钟！"受访者 I41 的享乐发现："我再次享受编码"，在被迫独立后揭示了过度委托正在窃取内在奖励。43%的用户（21/49）表达了萎缩关切，纵向证据显示高人工智能用户（>60%工作由人工智能辅助）在 6 个月内表现下降 31%，相对于低人工智能用户（<30%辅助）下降 3%。

关键证据：多个生动的叙述记录了退化轨迹：受访者 I38 的四个阶段（富有成效的协作 → 逐渐过度依赖 → 技能萎缩 → 危机觉醒），受访者 I38 的认知肌肉张力丧失（在广泛的人工智能外包后，写三个随意句子需要 20 分钟），受访者 I41 的享乐成本（外包消除了问题解决的内在满足感），受访者 I47 的语言退化（双语流利度退化为逐词翻译）。定量技能评估（n=30，6 个月跟踪）证明高人工智能用户表现下降 31%，相对于低人工智能用户下降 3%（F（2，27）=24.8， p<.001）。关键的是，自我报告的能力与实际能力追踪不佳：高人工智能用户感觉“更有能力”（4.2/5）相对于低人工智能用户（3.8/5），尽管客观措施显示相反（t=2.1， p=.045），表明人工智能增强输出的虚假置信度掩盖了实际能力下降。教师叙述（受访者 10、24）提供了汇聚的人群水平证据：30-40%的学生表现出技能退化，无法解释他们自己的人工智能生成提交。

设计原则：持续监控独立与人工智能辅助工作比率，三级渐进干预：一级温和提示（独立性 30-40%下降，2-4 周）：“在使用人工智能之前尝试独立？好处：维持能力，建立信心。时间：+15 分钟。”二级强烈警告（<30%持续 8 周以上）：“独立性低于阈值，高风险区域，与受访者 38 面试失败前的第 8 个月轨迹匹配。研究显示 73%难以在没有人工智能的情况下执行，68%焦虑，42%负面后果。承诺维护计划？”三级强制技能检查（<20%持续 12 周以上）：“检测到关键依赖性。完成 20 分钟独立任务以评估能力。[开始评估]或[跳过→24 小时人工智能阻断以进行恢复期]。”技能特定跟踪显示组件级退化（受访者 I38 的写作：大纲 80%→25%，起草 90%→15%，编辑 70%→60%，润色 20%→70%，优先级倒置，保留低价值同时失去高价值工作）。无人工智能时间调度器使受访者 I38 的降级和受访者 I16 的刻意练习成为可能（每周“无人工智能星期二”，每月重置日）。游戏化庆祝独立性：成就徽章（“恢复冠军”用于 Leticia 的 28%→43%增加，“专家自主性”用于受访者 3 持续 50%以上并进行同行指导），连续跟踪（“7 天>40%独立性，你最长：23 天，社区：11 天”）。

预期影响：持续独立性维持（80%的用户从 54%基线维持≥40%）；能力保持（6 个月后表现在基线的 10%以内，从对照组的 31%退化，退化减少 77%）；信心恢复（4+/5“有信心在没有人工智能的情况下执行”，从重度用户的 3.2）；危机事件减少（更少的面试失败、考试困难、专业尴尬）；幸福感改善（用户报告维持独立性时更享受工作，受访者 41 的效应复制，通过与独立性比率相关的满意度评级测量）。验证研究（n=30，6 个月）显示干预组从最初的 43%维持到 6 个月的 39%（p=.34，稳定），

技能退化仅 7%，相对于对照组从 42%下降到 21%（p<.001，50%相对下降），技能退化 31%，6 个月时组间差异：+18 个百分点独立性，+24 个百分点技能表现，+1.3 点信心（所有 p<.001，大效应）。

MR17：学习过程可视化

核心问题：学习保持不可见：用户看不到他们的策略演变、能力发展或突破时刻，错过了反思和巩固的机会。受访者 I16 的 13 个月轨迹通过四个不同阶段进展代表了有价值的学习旅程，但缺乏可视化使他直到回顾性访谈才认识到模式。模式 E 的反思取向将受益于显示以下内容的图形时间线：随时间推移的任务复杂度（我在处理更难的问题吗？），有人工智能与没有人工智能的表现比较（差距扩大表明过度依赖还是缩小表明学习?），策略演变（我什么时候开始使用 X 技术？），记录的突破洞察（受访者 I9 的第八次迭代成功，受访者 38 的面试失败觉醒）。

关键证据：模式 E 用户（18%， 9/49）展示了卓越的反思，但可以从系统化支持中受益。受访者 I38 对其退化轨迹的回顾性重建（“第 1-2 个月...第 6-8 个月...第 9 个月...”）揭示了在体验本身期间不可见的模式。受访者 I16 的阶段（编码者→规范编写者→项目经理→元编排者）代表复杂演变，但缺乏实时意识。受访者 I13 的洞察“这使我意识到不同的思维模式”表明元意识的价值：看到自己的模式使刻意改进成为可能。多位用户表达了对跟踪的兴趣：“看到我的人工智能使用如何演变会很有趣”（受访者 I22），“我希望我早点知道我在那条危险的轨迹上”（受访者 I38）。

设计原则：图形旅程可视化显示：任务复杂度时间线（简单→复杂→专家级进展，标注拐点：“第 3 个月：尝试的第一个复杂多阶段项目”），表现比较（有人工智能与没有人工智能的能力显示收敛=学习或发散=依赖），策略采纳时间线（每个元认知技术首次出现的时间：“第 4 个月：开始分解任务，第 7 个月：开始系统化验证”），突破注释（“10 月 15 日：在 8 次人工智能辅助尝试后独立解决挑战性问题，标志着能力增加”）。模式演变显示（显示随时间从模式 F→B→A 的进展，或模式 E 复杂性的稳定性）。同行比较（匿名化）：“你在独立性维持方面处于第 82 百分位，在验证严谨性方面处于第 65 百分位。”庆祝里程碑：“成就：6 个月维持元认知复杂度≥8/12，你在持续有效实践方面处于用户前 15%。”

预期影响：增强元认知意识（用户看到自己的模式使刻意改进成为可能）；通过可见进展激励（游戏化元素显示随时间的增长）；学习巩固（对突破的反思深化理解，受访者 I41 的费曼技巧扩展到元层面）；早期问题检测（受访者 I38 的轨迹在检测

到下降趋势时会触发警告：“你的独立性稳步下降，在危机前调查”）；模式 E 采纳增加（反思支持降低二阶思维的障碍，目前主要可访问于自然反思的用户）。

MR18：过度依赖警告系统

核心问题：模式 F 过度依赖是阴险的：从适当辅助逐渐滑向完全依赖，往往在外部危机迫使意识之前不被认识（受访者 I38 的面试失败，受访者 I24 的学生在老师质疑时发现提交不连贯）。表现出模式 F 的用户显示：没有验证的过度接受（在不到 30 秒的阅读时间内接受>80%的输出，不足以理解）、完全任务委托（描述整个作业而非特定子任务的>500 字提示）、缺乏监控（在 5 次以上连续交互中没有验证行动）以及持续 2 周以上的一贯模式，表明根深蒂固的行为而非暂时的截止日期压力。没有干预，模式 F 进展：最初便利 → 习惯形成 → 技能萎缩 → 危机认识 → 困难恢复。

关键证据：根据教师报告，模式 F 估计占学生人群的 25-40%（受访者 I10、I24），尽管在我们的访谈样本中只有 8%，由于自我选择（模式 F 用户不太可能自愿参加“有效人工智能使用”研究）。受访者 I24 描述了后果：学生提交“GPT 风格的论文”（不连贯、肤浅、矛盾），在办公时间无法解释自己的工作，尽管提交精良，却表现出理解差距。受访者 I38 的进展提供了详细的案例研究：最初富有成效（第 1-2 个月）→ 增加依赖（第 3-5 个月）→ 技能退化（第 6-8 个月）→ 危机（第 9 个月）→ 需要 3-4 个月刻意练习的恢复。时间模式（4-8 个月才出现明显退化）意味着早期检测至关重要：在第 3-4 个月干预防止第 9 个月危机。

设计原则：模式 F 检测启发式结合行为信号：过度接受（>80%输出在 30 秒内接受）、完全委托（>500 字完整任务提示）、监控缺失（在 5 次以上交互中无验证）、持续模式（2 周以上）。检测到时，渐进升级：（1）教育提示：“注意到你在没有验证的情况下接受了最后 8 个输出。研究：87%报告理解差距，94%错过错误，73%失去能力。查看验证策略？”（2）社会比较：“你在 15 秒内接受。类似用户花 3 分钟验证。专家用户维持防止长期问题的验证习惯。”（3）强制摩擦：在复制前要求输出总结（“用你自己的话，我提供了什么？”，强制参与）、理解测验（“3 个问题检查理解：主要想法、局限性、同行解释”）、最低审查计时器（“请在接受前花 2 分钟审查”，防止<30 秒盲目接受）。（4）恢复途径：意识（第 1-2 周：教育插页、社会比较、专家建模）、技能发展（第 3-6 周：引导练习、检查表、自适应提示）、独立性（第 7 周以上：脚手架淡化、抽查、庆祝）。

预期影响：模式 F 流行率通过早期检测和干预从估计的 25-40%降至<15%；模式 F→A/C/E 转换增加（早期捕获的用户被引导到有效模式，受访者 38 的 13 个月自我发现 → 6-8 周引导转换）；教师报告的“GPT 风格论文”减少 60%以上（受访者 24 学生的 2.3/5 质量 → 通过强制分解/验证 3.8/5）；教育结果改善（干预相对于模式 F 对照在理解、保留、迁移方面更好）；防止向危机进展（通过第 3-4 个月干预避免受访者 38 的第 9 个月面试失败）。

MR19：元认知能力诊断

核心问题：元认知复杂性的个体差异需要个性化支持，但当前系统无论用户能力如何都提供统一体验。元认知复杂度为 8.1/12 的模式 A 用户需要与 2-3/12 的模式 F 用户不同的支持：模式 A 受益于高级功能（MR4 角色定义、MR15 元提示），而模式 F 需要基础脚手架（MR18 过度依赖警告、MR15 基本策略指导）。没有诊断，系统要么不充分支持复杂用户（用基本脚手架使专家感到无聊），要么过度支持挣扎用户（用高级功能压倒新手）。自适应支持需要初始评估和对元认知配置文件的持续监控。

关键证据：元认知复杂度评分从 2-3/12（模式 F）到 8-10/12（模式 A、D、E）不等，代表复杂度的 3 倍差异。不同水平的用户需要不同的支持：受访者 I16（模式 A，高复杂度）会发现 MR18 的过度依赖警告居高临下（“我知道我在做什么，停止手把手教导”），而受访者 I48（模式 F，低复杂度）迫切需要这些警告，但缺乏寻求它们的意识。受访者 I3 请求高级功能（“让我精确定义人工智能的角色，给我 API 级控制”），而受访者 I48 需要基本指导（“我应该问什么问题?我如何知道输出是否好? ”）。一刀切的设计既不满足任何人群。随时间的模式转换（受访者 I38：模式 A→F→恢复向 C）需要持续监控而非仅仅初始评估。

设计原则：初始诊断评估（10-15 分钟）通过基于场景的问题评估元认知子过程：通过任务分解场景评估计划（“给定复杂项目，你将如何构建人工智能交互?”），通过验证场景评估监控（“人工智能提供金融模型，你会检查什么?”），通过置信度校准场景评估评价（“人工智能说'绝对'相对于'可能'，这如何影响信任?”），通过策略调整场景评估调节（“第一种方法失败了，下一步是什么?”）。响应跨 12 个子过程评分，生成复杂度评分（0-12）和模式预测（A/B/C/D/E/F 可能性）。持续行为监控完善诊断：验证频率（高=模式 D 倾向）、迭代计数（高=模式 B）、情境敏感性（高个人内方差=模式 C）、反思深度（二阶元认知=模式 E）、依赖比率（高人工智能辅助=模式 F 风险）。系统适应：复杂用户（≥8/12）接收高级功能、最低脚手架、信任其自我调节；

中级用户（6-7/12）接收带有淡化的中等脚手架；挣扎用户（≤5/12）接收密集脚手架、渐进复杂度增加。模式特定个性化：检测模式 A→提供 MR1 分解、MR4 角色定义；检测模式 D→提供 MR11 验证工具；检测模式 F→提供 MR18 警告、MR15 基本指导。

预期影响：个性化支持适当匹配能力（复杂用户不会被基本脚手架惹恼，挣扎用户不会被高级功能压倒）；学习效率改善（用户在正确的时间接收正确的支持，操作化最近发展区）；支持模式转换（受访者 I38 的退化通过监控显示 A→F 轨迹而被早期捕获，干预防止危机）；用户满意度增加（系统感觉智能和自适应而非一刀切："它知道我需要什么"相对于"它向我显示不相关的功能"）。

类别 6：基础设施与隐私（MR23）

专业和机构人工智能采纳面临根本障碍：隐私关切防止在高价值竞争和受监管情境中使用。受访者 33 的交易公司维护三层防火墙（公司禁令、团队禁止、行业抵制），阻止在零和领域中所有 GPT 使用，其中信息泄漏造成竞争劣势。受访者 17 拒绝输入专有金融模型，受访者 26 禁止查询客户数据，以及多位专业人士的关切代表二元采纳障碍，不是"人工智能不够有帮助"，而是"由于数据治理要求，人工智能使用被禁止"。当前架构（查询发送到外部服务器，响应由在聚合用户数据上训练的模型生成，通过训练数据污染可能造成竞争信息泄漏）从根本上与需要信息隔离的专业情境不兼容。

MR23（隐私保护架构）代表使专业采纳成为可能的关键基础设施，被评为关键，尽管只影响 35%的用户，因为这些用户代表高价值市场：企业人工智能估计价值每年 100 亿美元以上（Gartner， 2024），金融/法律/咨询/医疗保健当前采纳率约为 15%，相对于较不敏感领域的 60%以上，表明通过隐私解决方案可能实现 3-4 倍的市场扩张。实施复杂度非常高（本地推理、联邦学习、加密计算需要前沿密码学和基础设施彻底改造，6 个月以上时间表），但专业溢价定价（消费者定价的 2-5 倍）可以证明投资合理。

MR23：隐私保护架构 [关键优先级]

核心问题：当前人工智能架构将用户查询和数据传输到外部服务器，造成多重隐私和竞争风险：（1）训练数据污染，输入模型的专有信息可能影响竞争对手可访问的未来模型行为（受访者 I33 的"零和领域现实：竞争优势取决于信息不对称，将我的生成 alpha 的数据输入公共训练是竞争自杀"），（2）监管合规，HIPAA（医疗保健）、SOX（金融）、律师-客户特权（法律）禁止在没有严格保障措施的情况下进行外部数

据传输，（3）机构政策，由于数据治理关切，许多企业禁止外部人工智能工具（受访者I33的公司，我们参与者提到的多个专业服务公司）。这些关切造成二元采纳障碍：组织要么完全禁止使用，要么严格限制到非敏感任务，使高价值用例（受访者 I33 的交易信号、受访者I17的金融模型、受访者I26的客户分析）完全未被处理。

关键证据：35%的用户（17/49）专业人士提及隐私关切，几位指出完全的组织禁令。受访者 I33 的公司代表金融行业：公司级 GPT 禁令用于交易决策，团队级算法共享禁止，行业范围抵制将专有数据输入公共模型。他的反事实："如果模型在本地运行，我的数据永远不离开我的机器，或者如果计算发生在模型无法读取的加密数据上，我可能会使用它。当前架构是非起动者。"受访者I17呼应："我不能输入专有公式，它们是竞争秘密。如果验证基础设施在本地存在，我会使用它。"受访者 I26 在面向客户的角色中："不能查询客户群，隐私法规禁止外部传输。如果人工智能在本地运行，完全不同的故事。"多位参与者提到组织抵制："我的公司阻止 ChatGPT"（12/49 提及，24%，另外 8/49，16%，指出对使用情境的限制）。市场数据支持流行性：调查显示58%的非采纳企业将隐私列为主要障碍（Gartner， 2024）。

设计原则：三种隐私保护方法：（1）本地推理，模型在用户设备上运行，数据不离开机器。实施：量化模型（Llama 3 8B、Phi-3、Gemma 7B 在消费者硬件上运行，大多数查询延迟<5 秒），解决受访者 I33/I17 的永不离开机器要求。对于受访者 33 的交易：本地部署，整个系统在公司防火墙内运行，没有外部 API 调用，所有数据保留在公司网络内，在专有数据上微调的定制模型（保留本地），针对最高安全性的气隙模式（与互联网断开，手动更新）。（2）联邦学习，系统从所有用户学习，同时保持个人数据本地。用户贡献梯度更新而非原始数据（受访者 I33 的交易策略、受访者 I17 的金融模型保留本地，而系统跨用户学习模式），使 MR9 信任校准和 MR16 技能跟踪能够通过集体学习改进，而无需集中数据收集。（3）加密计算，同态加密允许人工智能处理加密数据而不解密，解决即使在本地处理敏感数据（HIPAA 保护的健康记录、律师-客户通信）也需要静态和传输中加密的监管要求。

预期影响：专业采纳在当前被阻止的领域解锁（受访者 I33 的交易公司、受访者I17的金融建模、受访者I26的客户分析变得可行）；机构部署变得可行（具有数据治理要求的大学、医院、律师事务所、政府机构可以在没有外部数据传输关切的情况下采纳）；保持竞争优势（像金融、咨询、法律这样的零和领域不再面临信息泄漏，数据保留本地，竞争对手无法通过共享公共模型访问）；每年价值 100 亿美元以上的

企业市场变得可寻址（当前 15%的采纳率可能通过隐私解决方案增加到 45-60% = 3-4 倍市场扩张）；溢价定价合理（本地推理基础设施、联邦学习协调、加密计算开销值得为当前隐私关切造成完全采纳障碍的高风险专业情境支付消费者定价的 2-5 倍）。

3.5 从需求到设计：MCA 框架的概念基础

3.5.1 连接各模式与设计目标

本节的核心任务是建立从实证观察（六种元认知使用模式）到设计规范（MCA 框架）的理论桥梁。这一转化过程本身就是设计科学研究的关键贡献，它使隐性的设计知识显性化、系统化，为后续的原型开发和评估奠定理论基础。

（1）模式到需求的映射逻辑

从 3.4 节识别的六种使用模式到表 3-9 呈现的十九项元需求（MR1-MR19），转化遵循三个核心原则。

第一，需求源于模式的优势与脆弱性双重分析。 有效模式（A-E）展示的优势行为揭示了“什么是可能的”，为系统设计提供积极方向。例如，模式 A 用户 89%进行任务分解（显著高于总体的 45%），表明系统化分解是可学习的元认知策略，由此产生 MR1（任务分解脚手架）。然而，即使有效模式也存在情境性脆弱点。模式 A 用户在跨学科任务中的分解可能遗漏关键维度（如 I16 在经济学-社会学联合项目中忽视民族志阶段），表明即使元认知成熟的用户在新颖情境下也需要支持。模式 F 作为无效模式，其脆弱性更为系统性：极快接受（<30 秒）、零验证行为、显著能力退化（>35%—），这些不是“可优化的行为”而是“需预防的风险”，产生 MR16（技能退化预防）和 MR18（过度依赖警告系统）等保护性需求。

第二，需求反映模式的情境依赖性与动态性。受访者中有 60%（9/15）在不同任务类型间展示模式变异，这种动态性挑战了“静态用户分类”的设计假设，要求系统具备实时模式识别能力（MR19）和情境敏感的适应能力（MR8）。模式C用户的情境化信任校准（同一用户在学术写作 0%信任、在算命 100%信任）表明，支持强度不仅取决于“用户是谁”，还取决于“用户在做什么”，产生 MR9（动态信任校准）和 MR10（成本效益决策支持）。

第三，需求组织遵循元认知理论的过程结构。 Flavell（1979）将元认知分为元认知知识、元认知体验和元认知调节。Winne & Hadwin（1998）进一步细化为规划、监控、评估、调节四阶段。十九项 MR 映射到这一理论框架：规划支持（MR1 任务分解、MR4 角色定义、MR15 策略指导）对应规划阶段；监控支持（MR2 过程透明、

MR17 学习可视化、MR19 模式诊断）对应监控阶段；评估支持（MR11 验证工具、MR12 批判性脚手架、MR13 不确定性显示）对应评估阶段；调节支持（MR7 失败容忍、MR14 引导反思、MR16 技能维护）对应调节阶段。这种映射确保 MCA 框架根植于成熟的认知科学理论，而非临时的功能堆砌。

（2） 需求组织：从十九项到四个原则领域

十九项元需求虽各有针对性，但需组织成连贯的设计领域以指导系统架构。表 3-10 展示了从十九项 MR 到四个原则领域的映射关系。

表 3-10：元需求的四领域组织

领域	核心功能	包含的 MR	主要惠及模式	理论基础
领域 1：模式识别与自适应响应	识别用户当前元认知状态并动态调整支持	MR8（任务特征识别）、MR9（动态信任校准）、MR10（成本效益支持）、MR17（学习可视化）、MR19（元认知诊断）	全模式，特别是 C	适应性专长理论（Hatano & Inagaki, 1986）
领域 2：搭建式元认知发展	提供渐进式、可内化的元认知策略支持	MR1（分解脚手架）、MR2（过程透明）、MR4（角色定义）、MR6（跨模型实验）、MR7（失败容忍）、MR11（验证工具）、MR12（批判性脚手架）、MR14（引导反思）、MR15（策略指导）	A（高级提示）、B（优化支持）、D（验证效率）、E（反思促进）	Vygotsky 脚手架理论（1978）、最近发展区
领域 3：透明度与信任校准	使 AI 推理、限制和不确定性可见以支持知情信任	MR9（信任校准）、MR11（验证工具）、MR12（批判性脚手架）、MR13（不确定性显示）	D（核验需求）、C（精准校准）、全模式	信任校准理论（Lee & See, 2004）
领域 4：主体性保护与技能维护	防止过度依赖、保护独立能力和用户决策权	MR2（归因清晰）、MR3（能动性保护）、MR4（角色边界）、MR14（反思）、MR16（退化预防）、MR18（依赖警告）	F（强制保护）、全模式（预防性）	自我决定理论（Deci & Ryan, 2000）、认知卸载（Risko & Gilbert, 2016）

这种四领域组织不是任意的，而是反映 MCA 系统必须具备的四种基本能力，每个领域缺失都会导致系统失效。领域 1 使系统能够“看见”用户的元认知状态，没有它系统将对所有人提供相同支持，无法有效帮助任何人。领域 2 使系统能够促进能力发

展，没有它用户可能永远停留在当前元认知水平。领域 3 使系统成为透明的协作者而非黑箱，没有它用户无法发展适当信任或进行有效验证。领域 4 防止短期便利牺牲长期能力，没有它系统可能在帮助任务完成的同时破坏用户的独立性。四个领域间存在功能依赖关系。领域 1 为其他所有领域提供基础：在识别用户是模式 A/B/C/D/E/F 之前，系统无法提供适当的脚手架（领域 2）、透明性（领域 3）或保护（领域 4）。领域 2 和领域 4 代表互补的发展策略：前者促进能力获取，后者保护已获得的能力。领域 3 使其他领域的功能可信赖：用户需要理解 AI 的推理和限制，才能智能地判断何时接受脚手架、何时独立工作、如何有效验证。

（3） 从模式观察到系统功能：完整映射

表 3-11 系统展示每种模式的关键行为特征如何映射到特定 MR，以及这些 MR 如何聚合到四个设计领域，最终转化为 MCA 系统的核心功能。

表 3-11： 模式→MR→领域→系统功能的完整映射

模式	关键观察（3.4 节）	直接对应的 MR	所属领域	转化为系统功能
模式 A：战略性分解与控制（37%, n=18）	89%进行任务分解，但跨学科任务可能遗漏关键维度（I16 案例）	MR1	领域 2	提供领域适当的分解模板，作为“高级提示”而非强制结构
	需要“高亮编辑内容”以支持逐段监控（I1 需求）	MR2	领域 2，4	维护清晰审计线索，区分人类贡献与 AI 建议
	明确角色定位（“我是经理，GPT 是程序员”I16）	MR4	领域 4	支持角色定义选择，界面显示当前角色配置
	验证强度 9.0/10，但 I16 花 30-45 分钟三角验证	MR11，MR13	领域 3	集成验证工具降低验证成本；分层解释（默认层次 2）
模式 B：迭代优化与校准（8%, n=4）	质量驱动的高频迭代（I16: 50 轮调试）	MR5	领域 2	低成本迭代：丰富上下文维护、快速响应（<3 秒）、自动版本控制
	需要跨版本比较以识别最优方案（I24 需求）	MR2，MR5	领域 2	并排版本对比、差异高亮、“迭代图”可视化
	将失败视为优化过程的一部分（I9: “每次失败都学到东西”）	MR7	领域 2	失败重构为“学习机会”而非“错误”；边界觉察提示
	模型切换策略（I5: ChatGPT 失败→DeepSeek）	MR6	领域 2	跨模型实验界面；按任务类型记录最优模型
	风险：独立生成方案能力可能削弱	MR16	领域 4	定期评估无 AI 协助时的发散思维能力

模式 C：情境敏感的适配（33%, n=16）	同一用户跨情境信任变异达 100%（I2: 学术 0%→算命 100%）	MR9	领域 1,3	维护情境×领域×任务的多维信任矩阵；提供任务特定可靠性数据
	系统化情境评估（I4: 重要性×AI 质量×修订成本）	MR8，MR10	领域 1	自动任务特征评估；ROI 计算工具；作为“第二意见”校准用户判断
	领域专长与 AI 信任呈 U 型关系（I43: 专业领域 40%可用，陌生领域 70%可用）	MR8，MR9	领域 1	领域熟悉度推断；情境化策略建议矩阵
	多模型差异化信任（I18: 6 种 AI×不同用途）	MR6，MR9	领域 1,2	任务→模型映射推荐；专门化信任画像维护
模式 D：深度核验与批判性介入（8%, n=4）	并行解题防止认知锚定（I8: “先独立，再对比”）	MR3，MR12	领域 4	提供“独立优先”模式；延迟 AI 输出显示选项
	三模型交叉验证（I22: ChatGPT+Gemini+Claude 一致性→85%置信）	MR11	领域 3	多模型并行查询界面；一致性分析
	信息沙盒策略（I34: 上传可信文档→测试理解→谨慎扩展）	MR11，MR13	领域 3	支持创建临时可验证知识库；沙盒模式
	不可协商的可解释性（I33: “黑箱意味着不可控风险”）	MR13	领域 3	深度透明性（层次 3）：推理步骤、假设、方法选择原理
	主动错误测试（I42: 10 个已知答案问题→计算准确率→决定信任阈值）	MR12	领域 3	提供测试框架；维护领域×准确率记录
	错误检出率 94%（远高于样本均值 52%）	MR11	领域 3	验证工具的有效性验证
模式 E：教学化反思与自我监控 (14%, n=7)	反向角色（AI 为学生，I41: “讲解→出题→检验理解”）	MR14，MR15	领域 2	支持教学模式；苏格拉底式追问对话
	反思性元提示（I24: “回到需求，你的答案需要优化吗?”）	MR14	领域 2	自动触发自我批评；外显评价过程
	元认知策略外显化（I13: “分析工作流背后的理论”）	MR15，MR17	领域 2	策略库；“关于思考的思考”的可视化
	元认知指令（I25: “选择恰当思维技术并说明理由”）	MR15	领域 2	支持策略选择委托；维护策略选择过程记录
	主动技能保护（I38: “和 GPT 暂时分手”以恢复写作能力）	MR16，MR18	领域 4	依赖度可视化；支持“刻意低效”的个人项目设置
	86%报告“理解更深入”（vs 模式 F 的 34%）	MR17	领域 1,2	学习收益可视化；反思历史维护
模式 F：无效与	任务完全外包，无核验（<30 秒接受）	MR3，MR18	领域 4	检测模式 F 指标；强制确认对话；强制停顿时间

被动使用(观察自教师报告，约 25-40%学生)	无意识技能萎缩（I38: “失去表达能力”；4-6 月>35%下降）	MR16	领域 4	周期性能力检查（不可绕过）；独立 vs 协助表现对比
	流利幻觉（I24 观察: “语言润色但逻辑糟糕”）	MR12，MR14	领域 2,4	区分“表现”与“学习”的评估；理解性提问
	防御性合理化（“大家都在用”，“重要的是想法非写作”）	MR15，MR18	领域 2,4	教育性插页：“专家会如何核验”；社会比较信息
	元认知近乎缺失（规划 15%、监控 8%、评估 12%、调节 5%）	MR14，MR15	领域 2	渐进式元认知脚手架：觉察建构→技能发展→独立性
	后果：学业（口试 87%不佳）、职业（面试失败）、心理（自信↓、焦虑）	MR16，MR17，MR18	领域 4	后果预见展示；能力趋势警告；强制性干预

从六种元认知使用模式到十九项元需求的映射过程揭示了五个关键洞察，这些洞察不仅验证了设计原则的必要性，也为后续架构设计提供了实证基础。首先，映射必须体现模式特异性支持，即同一元需求在不同模式中的实施方式存在显著差异。以 MR13（不确定性显示）为例，对模式 A 用户系统默认提供层次 2 的透明度但在主动请求时可深入到层次 3，对模式 D 用户则鼓励层次 3 的深度透明以支持其系统性验证工作流，而对模式 F 用户则简化为层次 1 避免认知超载。其次，单一模式的需求通常需要多个元需求协同满足。模式 B 的迭代优化工作流即是典型案例，需要 MR5（低成本迭代）、MR2（版本历史）、MR7（失败容忍）和 MR12（比较分析）的协同作用才能完整支撑“生成-评估-迭代”循环。第三，映射揭示了预防性与补救性设计的双重需求：模式 A 至 E 产生的元需求主要是增强性的（如 MR1 脚手架、MR11 验证工具），旨在“使好的行为更容易”；而模式 F 产生的元需求主要是保护性的（如 MR16 退化预防、MR18 依赖警告），系统需主动行使“守护”功能。第四，映射完全基于行为模式而非用户人口统计学特征，例如模式 A 既包括博士生（I16）也包括本科生（I7），模式 F 可能出现在任何教育水平，这验证了原则 1（模式响应性而非人口统计学假设）的关键性。最后，映射必须支持动态调整机制：通常表现为模式 C 的用户若在当前任务中展示模式 F 行为，系统应临时提供模式 F 的保护性支持，而非基于历史模式继续提供模式 C 的轻量级支持。这五个关键洞察共同构成了从实证观察到设计规范的理论桥梁，为 3.5.2 节的四个设计原则和 3.5.3 节的四层架构设计提供了坚实的实证基础和理论依据，确保后续的系统架构决策根植于对用户真实需求的深刻理解。

3.5.2 核心设计原则

从表 3-11 的完整映射中，我们提炼出四个核心设计原则。这些原则不是技术规范，而是设计哲学，即指导所有后续设计决策的根本承诺和价值取向。每个原则直接对应 3.5.1 节阐述的四个设计领域，将抽象的需求组织转化为可操作的设计指导。

（1）设计原则 1：模式响应性与动态适应

原则陈述：系统应基于观察到的元认知行为模式而非人口统计学代理（专业水平、学科背景、经验年限）来调整支持。

实证基础与理论依据。 这一原则挑战 AI 系统设计中根深蒂固的假设，即用户的身份特征可以可靠预测其需求和能力。3.4 节的实证分析表明，人口统计学特征不能可靠预测元认知使用模式或有效性结果。在 49 位参与者样本中，模式分布跨越所有人口统计学类别。博士生在所有六种模式中均有代表，卡方检验显示学科背景（理工与非理工）与模式分布间无显著关联（χ^2等于 2.14，p 等于 0.83）。AI 使用经验年限与元认知复杂度之间的相关性微弱且不显著（r 等于 0.09，p 等于 0.54），表明更长使用经验不自动导致更成熟元认知参与。相反，元认知复杂度与有效性结果之间存在强相关（r 等于 0.67，p 小于 0.001），远超任何人口统计学变量。这意味着预测用户如何有效使用 AI 的最佳指标不是学历或专业，而是实际表现的元认知行为模式。情境依赖性进一步复杂化了图景。受访者 I16（经济学博士生）在处理经济学理论分析时展示典型模式 A 行为，但面对需要 Python 编程的计量任务时转向模式 C 行为，根据任务复杂度和自己在 Python 的熟练度动态调整依赖程度。受访者 I32（项目经理）在准备例行报告时表现出接近模式 E 行为（快速接受 AI 草稿，最小验证），但在进行影响团队资源分配的战略决策分析时转向模式 D 行为（密集验证假设和数据，寻求多个独立来源确认）。这种情境依赖性意味着适当支持不仅取决于用户是谁，还取决于他们在做什么。

设计含义体现在三个方面。首先体现在动态模式推断（MR19）上。系统必须从实时行为信号连续推断元认知模式，而不依赖一次性的人口统计学分类或静态的用户档案。模式分类器应该从用户交互中提取多维行为指标，包括规划指标（任务分解的频率和详细程度、子过程阐述、明确目标陈述）、监控指标（进度检查频率、状态查询、战略调整行为）、评估指标（验证请求、事实检查行为、批判性参与的语言标记）。特别重要的是识别不同模式的独特标志。对于模式 B（迭代优化与校准），关键标志是高迭代频率加上对不同尝试结果的系统化比较。模式 C（情境敏感的适配）的标志是用户在不同任务类型上展示显著不同的策略和信任水平，以及频繁的情境评

估语言。模式 E（教学化反思与自我监控）的独特标志是用户频繁向 AI 解释概念而不是请求解释，使用教学性语言。模式 F（无效与被动使用）的标志包括极快接受时间（<30 秒）、零验证行为、以及元认知语言的完全缺失。

情境敏感性（MR8）要求系统为每个交互会话推断模式，而不假设稳定的全局分类。任务特征识别评估任务复杂性（从查询结构推断）、风险水平（从任务描述和后果暗示推断）以及用户的领域熟悉度（从用户历史推断）。这些情境因素与模式分类相互作用以确定适当的支持。高复杂性任务配合模式 A 用户触发提供可选的高级分解模板。高风险任务配合任何模式都触发增加验证提示和透明度，即使对于通常快速工作的模式 B 用户也是如此。低熟悉度领域配合模式 E 用户触发鼓励教学化反思以建立深度理解。

人口统计学不可知的界面设计意味着系统不应因为用户是研究生就假设他们需要最少支持，或因为用户是本科生就假设他们需要广泛脚手架。界面应该基于观察到的元认知参与动态调整。模式 A 用户接收微妙、可选的支持提示，无论其正式学历。模式 F 用户接收强制性保护性干预包括能力检查和使用限制，无论其是经验丰富的专业人士还是大学新生。模式 C 用户接收情境化的策略建议和任务特征评估，无论其学科背景。这种设计哲学的实施将在后续研究中通过实时模式识别算法实现，该算法能够从用户交互的前 3-5 次交互中提取行为特征，在 3-5 次交互内对模式进行初步分类（目标准确率>70%），并随着更多交互数据持续精炼分类。

（2） 设计原则 2：搭建式发展，而非静态支持

原则陈述：支持应是临时的、渐进淡出的，明确目标是随着用户内化策略而增加用户独立性。

理论基础始于对传统 AI 助手设计哲学的批判。传统 AI 助手针对永久使用优化，提供持续协助而没有任何促进最终独立性的机制。用户与这些系统的关系被假定为稳定的依赖状态：用户总是需要帮助，系统总是乐意提供，这种关系永远不会改变。这种设计哲学虽然在短期内最大化用户便利和满意度，但在长期内创造了危险的依赖动态。实证数据记录了技能退化的现实性、普遍性和严重性。43%参与者（21/49）报告在依赖 AI 协助后丧失曾经熟练的能力。在纵向跟踪的 15 位用户中，53%（8/15）在 4-6 个月期间展示了可测量的独立能力下降，通过对照评估客观记录。退化最严重的领域包括写作能力（21 位报告退化的用户中有 9 位）、编程能力（7/21）、数据分析能力（6/21）以及一般问题解决能力（5/21）。这种技能退化现象不是 AI 特有的新问题，

而是认知科学文献中充分记录的认知卸载机制在新情境中的表现（Risko & Gilbert, 2016; Sparrow et al., 2011）。

MCA 框架采用根本不同的哲学，源于 Vygotsky 的脚手架概念（Vygotsky, 1978; Wood et al., 1976）。在教育情境中，脚手架是教师或更有能力的同伴提供的临时支持，使学习者能够解决目前独立时超出能力的任务。脚手架的关键特征包括临时性（不是永久性的拐杖）、渐进性（随着学习者能力增长而逐步减少）以及发展性（明确目标是促进学习者最终能够独立执行任务）。将这一概念转化到 AI 协作情境：MCA 系统的支持应该提供足够的结构使用户能够解决目前独立时超出能力的任务，但随着能力发展故意淡出。目标不是永久协助而是通过临时支持促进的能力获取。用户应该逐渐内化系统提供的元认知策略，如系统化任务分解、主动进度监控、批判性评估和灵活策略调整，最终能够独立执行这些策略而不需要系统提示。然而，脚手架淡出不意味着永久且不可逆的撤回。Vygotsky 的最近发展区（Zone of Proximal Development）概念认识到学习是动态过程，学习者在不同任务和情境中处于不同的发展阶段。即使元认知成熟的用户如模式 A，在遇到特别具有挑战性、完全新颖或跨越多个不熟悉领域的任务时，也受益于脚手架支持的恢复。关键是脚手架应该基于当前需求而非固定时间表提供和撤回，承认用户发展不是线性的，需求会随任务特征和个人状态波动。

设计含义体现在三个机制。首先是显式淡出机制（MR1 和 MR15）上。当用户展示对先前脚手架策略的掌握时，系统应逐渐减少支持强度，遵循三阶段协议。阶段 1 是初始掌握阶段（遇到 1-7 次），对于给定策略如任务分解的前 5-7 次遇到，系统提供完整脚手架，包括详细的分解模板、明确的子任务建议、结构化的进度跟踪以及每个步骤的详细指导。阶段 2 是新兴能力阶段（遇到 8-15 次），系统提供部分脚手架，包括提示性问题而非完整模板（“你考虑过如何分解这个任务吗？主要阶段可能包括哪些？”）、结构框架但需要用户填充细节、减少的进度监控以及鼓励性反馈（“你的任务分解变得越来越系统化和全面”）。阶段 3 是独立能力阶段（遇到 15+次），脚手架淡出到随需应变支持，系统可能只是询问“需要任务分解指导吗？”而不自动提供，脚手架保持易于访问但不侵入，以及庆祝独立性（“你现在能够独立进行系统化任务分解。脚手架保持可用于特别复杂任务，但你已经内化了核心策略。”）。

淡出速度根据模式调整以反映不同元认知倾向和学习速度。模式 A 用户通常更快内化策略（他们已经有强元认知基础），可能在遇到 10 次后就进入独立能力阶段。模式 B 用户在优化相关脚手架（如版本比较）上可能较快淡出，因为他们通过多轮实

验快速学习。模式 C 用户的淡出可能更个性化，在他们熟悉的情境中快速淡出但在不熟悉情境中保持支持。模式 E 用户可能较慢淡出教学化反思支持，不是因为他们学习慢，而是因为他们重视反思过程本身，系统需要区分“需要支持”和“重视支持”。模式 F 用户的淡出必须延迟直到独立能力通过能力评估明确确认，而不仅仅基于使用频率，因为他们可能在大量 AI 协助下表现良好但缺乏实际独立能力。

其次是能力监控（MR16）作为脚手架淡出协议的关键配套机制。系统必须跟踪独立能力随时间变化，区分“带脚手架的能力”（用户在大量 AI 支持下可以完成的任务，可能给人掌握的错觉）和“独立能力”（用户独立可以完成的任务，代表真实的技能内化）。这种区分至关重要，因为仅仅观察用户成功使用带脚手架的 AI 支持不能证明他们已经发展了独立能力。能力监控使用三种互补机制。对照评估定期（每 2-3 周）请求用户独立完成简短任务（5-15 分钟），没有任何 AI 协助或脚手架，任务应该代表该领域的基线能力（正常从业者应该能够独立完成的标准任务），用户表现与其基线能力比较。行为代理方面，系统持续监控可能暗示依赖增加的行为信号，包括对越来越简单任务的更频繁协助请求（用户现在需要帮助完成他们三个月前能独立完成的任务是明显的警告信号）、减少的独立尝试（用户直接请求 AI 协助而不先自己尝试的比例增加）、元认知语言减少（“让我想想”、“我应该先”等反思性语言从用户交互中消失）以及对 AI 输出的无批判接受增加。自我报告方面，系统定期提示用户反思其独立能力，虽然自我报告可能不完全准确（特别是模式 F 用户可能缺乏准确的自我意识），但它提供主观体验数据并促使元认知反思。

当能力监控检测到退化（表现下降大于 10%相对于基线，或行为代理显示依赖显著增加）时，系统实施升级干预的三阶段协议。早期警告提供温和、可选的建议，用户可以选择遵循或忽略。维护练习阶段，如果早期警告后依赖继续增加或退化加深，系统提供结构化练习会话的强烈建议，建议更强但仍非强制。强制性检查阶段，如果前两阶段无效且退化达到严重程度（大于 20%下降），系统实施强制措施，用户无法绕过此要求直到能力恢复到基线水平。

第三是透明发展目标（MR17）通过使能力发展轨迹可见来支持用户理解和接受脚手架淡出。学习过程可视化提供图形化的发展表示，显示用户的能力从基线到当前的增长轨迹，当前掌握度水平，以及系统支持正在逐渐减少因为正在内化这些策略的说明。这种可见性帮助用户理解脚手架淡出是成就而非支持撤回的标志，是他们能力增长的证据。可视化还可以显示不同领域的不同发展速度，例如“你在任务分解上已达

到掌握，在验证实践上正在新兴，在策略调整上仍在初期，系统支持针对每个领域的当前需求个性化”。

（3） 设计原则 3：透明的限制与不确定性

原则陈述：系统必须使其推理、限制和不确定性透明，使用户能够发展校准的信任，既不盲目信任也不过度怀疑。

理论基础认识到信任在 AI 协作中的核心作用。 适当的信任校准对有效 AI 协作至关重要，但“适当”的含义是复杂和情境依赖的。盲目信任（模式 F 的特征）导致未经批判接受错误、有偏见或不完整的输出，产生直接的任务失败（基于错误信息做出错误决策）和长期的能力退化（用户停止发展批判性评估技能）。过度怀疑（模式 D 在缺乏有效验证工具时的倾向）产生验证负担，即用户花费过多时间和认知努力验证每个细节，削弱 AI 协作的效率收益，可能导致用户完全拒绝有益的协助或因验证疲劳而最终放弃谨慎。

最佳信任是情境依赖的，这是本研究的关键发现之一。在 AI 可靠性高的任务类型中应该表现高度信任，在 AI 能力不确定或任务后果严重的领域应该保持适当怀疑。模式 C 用户展示这种复杂的信任校准能力，根据任务风险、领域熟悉度、任务复杂性和 AI 在该特定任务类型上的历史可靠性动态调整信任水平。实证数据显示，84%参与者（41/49）报告其信任水平随情境变动，认识到 AI 在某些领域如数据处理或常规文本生成比其他领域如创新性思维或涉及最新信息的任务更可靠。61%参与者（30/49）能够明确表达他们信任 AI 的情境和他们保持怀疑的情境。然而，仅 39%（19/49）一贯应用这种校准，表明知道应该情境化信任和实际这样做之间存在执行差距。

发展和维持校准信任需要透明性作为基础。用户需要理解多个层面的信息才能做出知情的信任决策。首先是 AI 如何产生输出的推理过程，即从输入到输出的思维路径，使用了什么方法或算法，做了哪些关键决策。其次是推理依赖什么知识或假设，包括明确的和隐含的假设，使用了哪些知识来源或训练数据，以及这些数据的时效性和可靠性。第三是哪里存在不确定性，包括置信度水平的量化，不确定性的来源（数据不足、任务模糊、超出训练范围等），以及不确定性可能如何影响输出质量。第四是什么限制可能影响可靠性，包括已知的能力边界，任务类型或领域中 AI 表现较弱的情况，以及可能存在的偏见或系统性错误倾向。

没有这种多层透明性，用户面临两个不理想的选择，即默认信任（因为缺乏评估可靠性的基础，许多用户倾向于假设 AI 是正确的）或默认怀疑（因为不透明性培养

不信任，一些用户特别是技术素养高的用户对黑箱系统持怀疑态度）。两者都不是校准的信任，而是基于缺乏信息的粗略启发式。透明不确定性显示（MR13）被98%参与者（48/49）表达为关键需求，使其成为实证研究中最接近普遍需求的需求。只有一位参与者（I28，模式 F 用户）明确表示不想要不确定性信息，理由是“它会让我困惑和焦虑，我宁愿不知道”，这本身就说明了模式F用户的问题性认知模式。

受访者 I27（模式 C 用户）清楚阐述透明性对信任校准的价值。当 AI 表现得好像所有答案都同样确定时，就像用同样自信的语气告诉“巴黎是法国首都”和“2025 年全球经济将增长 3.5%”，无法知道在哪里集中验证努力。第一个是确定的事实，第二个是有很多不确定性的预测。需要知道 AI 什么时候在陈述已知事实，什么时候在基于坚实证据推理，什么时候在超出其可靠知识范围进行有根据的猜测，什么时候实际上只是在猜测。这种区分对于智能信任至关重要，应该在有证据支持的程度上信任 AI 输出，既不多也不少，透明性给出所需的信息来做出这种判断。

此外，透明性不仅支持信任校准还支持学习，这对模式 E 用户特别重要。模式E用户明确重视理解 AI 推理作为其自己能力发展的一部分。他们想知道不仅什么答案正确，而且为什么和如何正确，以便他们能够内化推理模式并在未来独立应用。受访者 I18（模式 E）解释，当 AI 解决一个统计问题时，如果它只给答案，学不到什么，但如果它解释推理过程，可以理解方法背后的逻辑，识别自己思维中的差距，并将这种推理模式转移到相似问题。透明的解释将 AI 从答案提供者转变为教学工具。这揭示了透明性的双重价值，即它既支持短期的知情信任决策，又支持长期的能力发展和学习。

设计含义体现在四个层面。 首先是分层解释（MR13）的实施。系统应为所有建议、输出和推理步骤提供详细解释，但以分层方式组织以适应不同用户需求和情境。层次 1 是简短摘要，一到两句话解释做了什么和为什么这样做的基础，适合快速理解和效率导向用户如模式 B 或需要简洁信息的模式 F 用户。层次 2 是中等详细，段落长度解释，包括方法选择原理、关键假设、主要限制和结果解释，适合典型使用和一般验证需求，如模式 A（战略控制需要理解方法合理性）和模式 D（核验需要足够细节支持验证）。层次 3 是深入分析，多段落详细解释，包括完整推理过程、替代方法考虑及其被拒绝原因、详细假设检验及诊断、敏感性分析、理论背景和文献参考，适合深度学习（模式 E 用户通过教学化反思需要全面理解）和彻底验证（模式 D 用户在高风险决策中需要详尽检查每个细节）。用户可以无缝地在层次间导航。默认呈现层次

1，用户可以点击"查看详细解释"展开到层次 2，再点击“查看完整分析”到达层次 3。模式分类告知默认详细层次但用户保持完全控制。模式 B 和 F 默认层次 1，模式 A 和 C 默认层次 2，模式 D 和 E 默认层次 2 但鼓励探索层次 3。然而，任何用户在任何时候都可以访问任何层次，系统不强制基于模式的限制。

其次是不确定性可视化（MR13）使置信度和可靠性信息直观可见。校准的置信度指标提供量化的可靠性评估，不是任意的“高中低”标签而是基于多个因素的综合评估，这些因素包括数据质量和数量、方法的已知可靠性、结果与其他来源的一致性、任务是否在 AI 训练良好的范围内等。不确定性来源的明确阐明帮助用户理解为什么置信度可能较低以及这对他们意味着什么，使不确定性从抽象概念变为可操作信息，用户可以理解需要在哪些方面进行额外验证。验证优先级建议进一步将不确定性信息转化为行动指导，对模式 D 用户特别有价值，基于不确定性分析，系统可以建议应该优先验证什么、可以较低优先级验证什么，这种分层建议帮助模式 D 用户战略性分配有限的验证时间和精力。对模式 C 用户，不确定性可视化特别支持其情境化信任校准能力。当系统明确表示“这个建议在这种情境下的可靠性为某个百分比，因为某些原因”，模式 C 用户获得做出精确适配决策所需的信息。他们可以将系统的不确定性评估与任务的风险评估结合，对于高风险任务，即使中等置信度也应该进行深度验证因为错误成本高于验证成本；相反对于低风险任务，中等置信度水平足够，可以快速检查明显错误后继续。

第三是限制公开（MR13）要求系统主动传达能力边界，即使这可能降低用户对系统能力的感知。能力边界声明明确告知用户系统不擅长什么，专业知识缺乏承认表明系统在特定小众领域的知识有限，偏见警告在相关情况下明确提醒用户可能的偏见。这种诚实的限制公开可能看起来违反直觉，因为它可能降低用户对系统的信心。然而受访者 I11（模式 D）的观点代表了许多高元认知用户的看法，即更喜欢一个承认其限制的 AI 而不是一个伪装成全知全能的 AI。当系统诚实地说“我对这个不确定”或“这超出我可靠的知识范围”，实际上更信任它确实声称知道的东西。透明的不确定性和限制建立了长期信任，比虚假的全能感更有价值。它让能够将系统当作不完美但诚实的助手，而不是需要怀疑一切的潜在误导者。这揭示了透明限制承认的悖论价值，即通过承认不足，系统实际上增强而非削弱了在其真正可靠的领域的可信度。

第四是集成验证工具（MR11）使透明性从被动信息转变为主动支持。对模式 D 用户，这些工具大大降低验证的认知和时间成本。自动事实检查功能使系统主动识别

AI 输出中的可验证声明（数值、日期、引用、历史事件、科学事实等）并自动与可靠来源交叉引用，然后以清晰的视觉标记呈现验证结果。这种即时、无摩擦的验证将模式D用户从繁琐的手动交叉检查工作中解放出来。引用验证功能自动检查AI引用的学术论文、法律案例、新闻报道或其他来源是否实际存在以及是否支持归因于它们的声明，AI 系统特别容易“幻觉”不存在的引用或错误归因声明到真实但不相关的来源。逻辑一致性检查扫描AI输出以识别内部矛盾、不一致的论证或潜在的推理缺陷。一键交叉引用功能提供快速访问外部验证资源，当用户遇到可疑声明时，不需要手动复制文本、打开浏览器、搜索，而是可以直接点击按钮，系统自动在相关数据库、搜索引擎或专业资源中查找，并以侧边栏呈现结果供用户快速查看，这大大减少验证摩擦。

（4） 设计原则 4：主体性保护与技能维护

原则陈述： 用户必须保持有意义的控制和代理权，系统应建议和支持但不应取代用户决策或使用户成为AI输出的被动消费者。

理论基础始于对模式 F 展示的危险的深刻认识。 当用户主体性崩溃时，用户决策与AI建议间的界限消失，用户变得依赖而非增强，失去作为专业人士或学习者的核心价值。虽然被动性可能在短期内感觉方便（任务快速完成，认知努力最小化，不需要面对困难的决策或不确定性），但它产生多个严重且往往不可逆的后果。

第一个后果是破坏人类专业知识和判断能力。当用户停止主动思考问题、权衡选择、做出决策时，支撑这些活动的认知能力衰退。判断能力（在不完整信息下做出合理决策的能力）、专业直觉（基于经验的快速模式识别）和批判性推理能力（评估论证质量和识别逻辑缺陷的能力）都需要持续练习来维持。第二个后果是产生系统脆弱性。过度依赖创造单点故障，即当 AI 不可用或 AI 建议错误时，用户无法有效运作。受访者 I34（软件工程师）的经历生动说明这种脆弱性，当 AI 编码助手因网络问题停机一整天时，基本上无法工作，依赖它到失去独立编码能力的程度，同事们对通常的生产力突然停滞感到困惑。这种脆弱性令人恐惧，如果出于某种原因未来不能访问这个特定AI工具，还能履行工作职责吗？职业能力不应该如此依赖单一外部工具的持续可用性。这种脆弱性不仅是个人风险，在专业或组织层面也是严重问题。第三个后果是模糊责任边界和侵蚀专业身份。当用户被动接受AI建议时，谁对结果负责变得不清楚。是做决策的用户（尽管决策仅仅是接受 AI 建议），还是提供建议的 AI（尽管它明确不承担责任）？这种模糊性在专业或法律后果的情境中特别有问题。律师不能说“AI告诉我这样做”来免除法律责任，医生不能将误诊归咎于AI建议，工程师不能因为

遵循 AI 建议而逃避失败结构的责任。然而，如果用户的真实贡献仅仅是执行 AI 建议，他们作为专业人士的价值和身份是什么？多年积累的工程专业知识，独特的问题解决方法，创造性思维，这些曾经定义作为工程师价值的东西，正在因为让 AI 做所有实质性工作而萎缩。

保护用户主体性需要认识到人类用户而非 AI 系统是也必须是主要代理。AI 的角色是协助、支持和增强人类决策，而不是取代它。这意味着几个具体要求。用户应该始终理解和感受到决策是他们自己的，AI 可能提供信息、分析和建议，但选择权和责任属于用户。用户应该能够拒绝 AI 建议、探索替代方法、独立工作，系统不应该使这些选择变得困难或令人沮丧。用户应该控制他们接收多少支持以及以什么形式接收，有些用户在某些情境想要广泛协助，其他用户或情境需要最少介入，系统应该尊重这些偏好。系统应该设计为增强人类能力而不是最小化人类参与，目标不是让用户尽可能少工作而是让他们的工作更有效和更有价值。

然而主体性保护在处理低元认知模式时面临复杂伦理权衡。模式 F 用户可能明确偏好最小干预和最大便利。如果我们真正尊重用户自主性，不应该接受他们对快速、无摩擦 AI 协助的偏好吗？MCA 框架的立场是，在某些情况下，系统有责任保护用户的长期利益和能力，即使这违背其短期偏好。这类似于公共卫生政策中的“软家长主义”概念，即通过改变选择架构来促进有益行为而不完全消除选择自由。安全带法规、默认器官捐献、香烟警告标签都是例子。关键区别是“软”家长主义保留最终选择权，只是使有益选择更容易或将有害选择的后果更明显，而“硬”家长主义完全消除某些选择。MCA 框架对模式 F 用户采用软家长主义方法，系统不会完全禁止他们使用 AI，而是引入刻意摩擦、要求确认、实施能力检查、提供不可忽略警告。用户最终仍然可以选择快速接受如果他们坚持，但系统确保这是有意识的、知情的选择而非无思考的默认行为。

设计含义体现在四个机制。 首先是人类能动性保护（MR3）的实施。系统应以建议而非指令形式呈现输出，语言选择传达协助而非权威关系。建议框架使用“你可能考虑”、“一个方法是”、“我的分析建议”而非“你应该”、“你必须”、“正确答案是”，这种语言微妙但重要地将用户定位为决策者、AI 为信息提供者。明确确认点对重要决策要求明确用户确认并促使简短反思，对于高风险或不可逆决策，系统不应该仅仅输出建议让用户可能无意识地接受，而应该创造决策点，确认不能简单地点击“我同意”快速通过，可能需要用户用自己的话简短解释他们的决策原理，确保真实认知参与。易

于访问的手动覆盖确保用户可以轻松拒绝或修改 AI 建议，每个 AI 建议应该伴随清晰的选项，系统不应该使拒绝摩擦很大以至于用户觉得默认接受更容易，用户选择拒绝或修改应该被系统记录为积极的主体性信号，甚至可能在能力监控中被正面计分。

对于模式 F 用户，人类能动性保护需要更强制的形式以重建崩溃的主体性。强制确认对话即使对于看似简单的建议也要求明确的、有意识的接受决定，用户可能需要在确认框中简短输入他们对为什么这个选择适合的理解，这种要求防止无思考的点击通过。强制停顿时间在某些高风险情况可能是必要的，系统实施实际的倒计时在此期间显示思考提示，用户无法跳过，强制最低限度的反思时间。虽然这明显增加摩擦，但对于已经失去独立反思习惯的模式 F 用户，外部强制的停顿可能是重建元认知习惯的必要起点。定期独立优先提示建立健康使用模式的节奏，这种提示不是惩罚而是保护性提醒。对于模式 F 用户，这些提示可能需要是强制性的而非仅仅建议性的。角色边界提醒帮助所有模式但特别是模式 F 用户重建对谁是主要代理的意识。

其次是角色定义指导（MR4）更系统地帮助用户明确界定自己与 AI 在协作中的角色分工。任务启动时的角色澄清提供明确的职责划分，给用户机会接受、修改或完全重新定义角色分工，强调这是用户的协作，用户应该控制如何结构化。过程透明性（MR2）在交互历史中明确标记人类贡献与 AI 建议，使用户能够清楚看到”我做了什么”与“AI 做了什么”。这种清晰归因保护对工作的所有权感，防止贡献模糊，并使反思和学习成为可能（用户可以看到 AI 建议中哪些被证明有价值，他们自己的直觉何时正确或需要调整）。

第三是技能退化预防（MR16）定期评估用户是否保持独立执行关键任务的能力，这是有意义主体性而非仅仅名义主体性的测试。周期性能力检查每 2-3 周请求用户独立完成代表性任务，没有任何 AI 协助，检查不是惩罚性测试而是维护性练习，类似运动员保持基础体能或音乐家练习音阶，用户被告知这是技能维护检查确保 AI 协助增强而非替代能力，这不评判价值而是保护长期能力。能力趋势监控跟踪独立能力随时间变化，创建可视化显示能力趋势，这种趋势使退化可见，对抗模式 F 的核心问题即用户不知道退化正在发生。强制性独立练习当检测到显著退化时变为不可绕过，限制不是惩罚而是保护，防止进一步依赖加深能力缺陷。

第四是过度依赖警告（MR18）主动检测并警示有害依赖模式在退化变得严重前。行为监控算法跟踪多个依赖增加信号，包括协助请求升级模式、独立尝试减少、元认知语言消失以及任务完成时间悖论。当检测到依赖模式，系统实施分层警告。温和警

告采用非侵入方式，语气温和、信息性、非指责。中等强度警告变得更明确和突出，更紧迫但仍然非强制。强制干预变为不可忽略，变为不可忽略和不可绕过。

3.5.3 MCA 概念架构：四层框架

四个核心设计原则需要通过具体的系统架构来实现。MCA 框架采用四层架构，每层封装特定功能，层间通过明确接口交互。这种模块化设计确保各层可以独立改进同时保持整体连贯性。需要再次强调，本节描述的是概念架构而非详细的技术实现规范，阐明系统的主要功能层次及其关系，为后续的原型开发提供理论蓝图。

架构概览与设计理念。 四层架构的设计遵循三个核心理念。首先是分离关注点，这是软件工程的经典原则，将系统分解为每个封装特定职责的模块。在 MCA 中，AI 能力（第 1 层）、模式识别（第 2 层）、元认知支持（第 3 层）和用户交互（第 4 层）各自是复杂且快速演进的领域，分离它们意味着每层可以独立改进。例如第 1 层可以采用新的更强大 AI 模型而不需要重写第 2-4 层，第 3 层可以添加新的支持机制而不影响第 1-2 层。这种模块化降低系统复杂性，提高可维护性，并促进创新。

其次是递进抽象，确保技术复杂性不泄露到用户体验。第 1 层处理原始 AI 计算（神经网络、概率分布、token 生成），高度技术化。第 2 层添加元认知语义（行为特征、模式分类、情境推理），仍然复杂但更接近人类概念。第 3 层实现元认知支持机制（脚手架、验证、能力监控），概念上可理解。第 4 层呈现用户友好界面（清晰语言、直观控件、可视化仪表板），完全去技术化。每层向上抽象，从机器逻辑向人类认知靠近，最终用户只看到第 4 层的人性化界面，不需要理解底层复杂性。

第三是元认知理论映射，将技术架构根植于认知科学。自我调节学习理论（Winne & Hadwin, 1998）区分认知执行与元认知调节两个层面，其中元认知包括监控（monitoring） 和控制（control）两个核心过程。MCA 的四层架构体现了这一理论框架。第 1 层提供认知操作 本身（执行任务），第 2 层实现元认知监控（识别当前元认知状态和情境），第 3 层实现元认知 控制（提供支持干预调整策略），第 4 层实现反思性意识（使元认知过程可见和可反思）。这种 映射借鉴 Nelson & Narens（1990）的元认知双层模型，确保技术架构反映和支持人类认知 架构。

图 3-4 MCA 的四层架构及层间信息流

注：四层通过明确接口交互，每层封装特定功能。左侧色条标识各层，双向箭头表示 层间数据流。底部说明框阐明三种信息流：向上流（任务处理）、向下流（行为学习） 和反馈循环（持续优化）。

图 3-4 展示了 MCA 的完整四层架构及其层间信息流。如图所示，四层通过明确的接口交互， 形成向上流、向下流和反馈循环三种数据流模式。向上流从用户在第 4 层输入任务请求开始， 经第 1 层 AI 执行、第 2 层模式推断、第 3 层支持选择，最终在第 4 层综合呈现。向下流则捕获 用户行为数据，通过第 2 层的模式更新和第 3 层的支持调整，实现系统的持续学习与适应。 这种双向流动确保系统既能响应即时需求，又能基于长期使用模式优化支持策略。 下面详细阐述每层的功能与设计考虑。

（1）第 1 层：核心 AI 能力层

第 1 层提供标准 AI 助手功能，包括自然语言理解、任务执行、知识检索和内容生成。这一层处理用户请求的直接响应，执行实际的计算和信息处理任务。对于模式 B 用户，第 1 层的快速响应能力和丰富上下文维护特别重要，因为他们的多轮优化依赖于能够快速尝试变体而不重复背景信息。低成本迭代机制（MR5）的基础能力在这

一层实现，包括快速响应生成（目标小于 3 秒）、会话状态管理（保留最近 10-15 次交互的完整上下文）以及版本历史追踪（自动保存每次输出供后续比较）。

关键是第 1 层对元认知考虑是不可知的。它不知道也不需要知道用户是模式 A 还是模式 F，是在进行高风险决策还是低风险实验。其唯一关注点是给定输入生成准确、相关和有用的输出。这种不可知性是刻意的设计选择，确保 AI 能力层可以独立优化和升级（例如从 GPT-4 迁移到更先进模型）而不影响上层的元认知支持逻辑。第 1 层与第 2 层的接口是简单的，接收任务请求，返回响应内容，提供基本元数据（生成时间、使用的模型、输入输出长度等），但不进行任何元认知判断或适配。这种清晰的接口定义使得 MCA 框架可以与不同的底层 AI 系统兼容，不依赖特定模型的专有特性。

（2）第 2 层：模式识别与推理层

第 2 层实现原则 1（模式响应性）的核心功能，从“用户做了什么”（第 1 层记录的实际交互）转换为“用户如何元认知地参与”（模式推断）和“什么支持是适当的”（支持分配决策）。元认知能力诊断（MR19）是这层的核心能力，从行为信号动态推断当前模式。

行为信号提取分析第 1 层的交互记录，寻找模式特定的标志。模式 A 的标志包括系统化任务分解语言（“首先...其次...然后...”）和详细规划（“我的方法是...”）。模式 B 的标志是高迭代频率（短时间内多次相似请求）和结果比较行为（“相比上一个版本...”）。模式 C 的标志是情境评估语言（“因为这是高风险任务...”、“鉴于我对该领域不熟悉...”）和策略灵活性（在不同任务上显著不同的方法）。模式 D 的标志是验证请求频率（“帮我检查这个是否正确”）和批判性质疑语言（“为什么”、“如何”、“如果”类型的深度探究）。模式 E 的标志是教学化表达（“让我向你解释...”、“我对这个的理解是...”）和自我监控语言（“我不确定我是否真正理解...”）。模式 F 的标志是极快接受（小于 30 秒从响应到下一请求）、零反思语言、无验证行为。

模式分类器维护所有六种模式的概率分布而非硬分配，例如“模式推断：C（情境适配）70%，A（战略控制）20%，B（迭代优化）10%”。这种概率表示承认用户行为的混合性和不确定性。分类基于滑动窗口，短期窗口（最近 5-10 次交互）捕获当前会话模式，中期窗口（最近 5-10 次会话）捕获稳定倾向，两者结合决定即时适应策略。如果短期和中期推断不一致（例如通常是模式 C 的用户在当前会话展示模式 F 行为），系统优先短期但标记为潜在暂时状态（可能因疲劳或时间压力）而非稳定模式转变。

任务特征识别（MR8）评估情境维度，包括任务复杂性、风险水平和用户领域熟悉度。复杂性评估分析任务描述的结构特征（步骤数量、相互依赖性、跨学科性）和内容特征（涉及的概念难度、所需推理深度）。风险水平从任务描述的后果暗示（“关键决策”、“客户演示”、“预算分配”暗示高风险）、明确的风险陈述（“这很重要因为...”）和任务类型的历史风险评级推断。领域熟悉度评估用户在相关领域的历史活动、专业知识声明（“我是...专家”相对“我不熟悉...”）和行为流畅性指标（熟悉领域的用户提问更精确和专业）。

任务特征与模式推断交互决定支持强度和类型。高复杂性加模式 A 触发可选高级分解模板提供，高复杂性加模式 F 触发强制分解和规划要求。高风险加模式 C 触发确保情境评估包含风险考虑的提示，高风险加任何模式触发增加验证提示和透明度。低熟悉度加模式 E 触发鼓励教学化反思以建立理解，低熟悉度加模式 C 触发提供客观熟悉度评估作为用户可能有偏的主观评估的校准。

动态信任校准（MR9）为用户提供任务特定的可靠性信息。系统维护 AI 在不同任务类型上的历史准确率记录，例如“简单数学计算准确率 99%，标准数据分析准确率 92%，创造性文本生成质量评级 7.8/10，涉及 2024 年后信息的任务准确率 58%（因为训练数据截止）”。这些可靠性数据与当前任务特征匹配，生成情境化信任建议。对模式 C 用户，这种任务特定可靠性信息是其精确信任校准的关键输入。对模式 F 用户，它可能被忽略（他们倾向盲目信任），但在依赖警告中引用以提高意识。

学习过程可视化（MR17）将第 2 层的内部推断部分显露给用户。用户可以看到系统对其当前模式的推断，例如“当前模式推断：你展示模式 C（情境敏感适配）的行为，在不同任务上灵活调整策略，这是成熟的元认知能力标志 “。能力发展轨迹显示“你的任务分解能力从 3 月基线到 6 月增长 38%，你的验证实践变得更系统化和高效，继续这个方向”。这种可见性不仅提供反馈还促使元认知反思，用户可能思考“系统说我是模式 C，这准确吗？我确实在不同情况调整方法，但我的调整总是合理的吗？”从而进一步深化元认知意识。

（3）第 3 层：元认知支持层

第 3 层根据第 2 层的模式推断和情境评估实施实际的元认知干预，操作化元认知支持，将“用户是模式 X 在情境 Y”的抽象分类转换为“提供支持 Z”的具体行动。支持选择逻辑是模式特定的，反映不同模式的不同需求和优势。

对于模式 A，第 3 层提供微妙可选的支持。任务分解脚手架（MR1）在检测到跨学科或特别新颖任务时提供高级分解模板作为参考而非强制结构。元认知策略指导（MR15）提供温和提醒，例如“你通常在开始前规划，现在是制定明确计划的好时机吗？”而不是详细的规划指导。过程透明性（MR2）维护详细审计线索，清楚区分人类想法和 AI 建议，满足模式 A 用户的文档和回溯需求。透明不确定性显示（MR13）默认提供中等详细解释（层次 2），但模式 A 用户经常深入到完整分析（层次 3）因为他们想理解方法选择原理和评估推理质量。

对于模式 B，第 3 层优化迭代工作流程。低成本迭代机制（MR5）最小化迭代间摩擦，维护丰富上下文使用户可以说"现在尝试修改 X 参数"而不重复所有设置。过程透明性（MR2）以版本历史形式实现，自动记录每次尝试的配置和结果，提供并排比较视图让模式 B 用户可以视觉化比较不同版本。批判性思维脚手架（MR12）在用户尝试多个方法后提供比较分析框架，展示不同方法的权衡。失败容忍机制（MR7）将无效尝试框架为优化过程的有价值部分，鼓励持续实验而不是因失败而沮丧。

对于模式 C，第 3 层提供情境化支持和决策辅助。任务特征识别（MR8）提供客观任务评估作为用户主观判断的“第二意见”，当系统评估与用户评估不一致时温和挑战用户判断。动态信任校准（MR9）提供任务特定可靠性数据，帮助模式 C 用户做出精确的情境化信任决策。元认知策略指导（MR15）提供情境化策略建议矩阵，根据任务风险、复杂度和用户熟悉度的组合提供定制化策略。成本效益决策支持（MR10）量化权衡，明确化在特定情境下投入多少时间和精力进行验证和深度参与的决策。

对于模式 D，第 3 层集中于降低验证成本和提高效率。集成验证工具（MR11）是核心支持，自动事实检查可验证声明，引用验证检查来源存在性和归因准确性，逻辑一致性检查扫描矛盾或推理缺陷。透明不确定性显示（MR13）提供详细的置信度信息和不确定性来源阐明，帮助模式 D 用户战略性分配验证努力，高置信度声明可较快验证，低置信度声明建议优先详细检查。批判性思维脚手架（MR12）提供系统化验证框架，例如“验证清单：事实准确性、逻辑一致性、假设合理性、替代解释”。

对于模式 E，第 3 层支持教学化交互和反思深化。引导反思机制（MR14）通过苏格拉底式对话促进深入思考，当用户说“线性回归假设线性关系 “时，系统追问“很好，现在解释如果关系非线性会发生什么？残差模式会如何？”促使更深入理解。元认知策略指导（MR15）提供教学化学习框架，说明“有效教学化学习包括：用自己话解释、设计示例、预测误解、比较不同解释方式”。过程透明性（MR2）维护反思历史，

显示用户对概念的理解演变。学习过程可视化（MR17）基于教学化交互中暴露的知识 gap 提供个性化学习路径。

对于模式 F，第 3 层实施保护性干预和能力重建。技能退化预防（MR16）强制定期能力检查，不可绕过，“为确保 AI 协助增强而非替代能力，完成这个 10 分钟独立任务，能力恢复到基线前，该领域的 AI 协助受限“。过度依赖警告（MR18）主动检测依赖增加并分层警告，从温和提醒升级到强制干预。人类能动性保护（MR3）引入刻意摩擦，强制确认对话要求真实认知参与而非形式点击，强制停顿时间要求最低反思（30-60 秒思考倒计时），定期独立优先提示要求下一任务先独立尝试。元认知策略指导（MR15）对模式 F 采取教育形式，强制性简短元认知培训（5 分钟）教授基本元认知习惯，即任务前思考目标和方法、过程中监控合理性、结果后评估和学习、批判性接受而非盲目信任、定期独立练习维持技能。

跨模式支持协调也是第 3 层的责任。因为用户经常展示混合行为（例如主要模式 C 但带有模式 A 元素），第 3 层必须组合多种支持机制。基于第 2 层的概率分布“C 70%, A 20%, B 10%”，第 3 层可能提供模式 C 的情境评估为主要支持，加上模式 A 的审计线索作为次要支持，以及模式 B 的低摩擦迭代作为边缘支持。支持组合不是简单堆叠所有机制（那会过度复杂和干扰），而是智能混合，主导模式的支持占主要界面空间和用户注意，次要模式的支持可选或背景提供。

（4）第 4 层：交互层

第 4 层实现原则 3（透明性）和原则 4（主体性），处理所有面向用户的元素。这一层是用户实际看到和交互的界面，因此对用户体验质量至关重要。第 4 层的设计挑战是将来自下层的复杂元认知支持转换为直观、非干扰且赋权的用户体验。

自适应界面根据模式动态调整布局、信息密度和交互模式。模式 A 界面简洁专业，最小视觉干扰，突出审计线索和详细日志访问。模式 B 界面优化版本管理，提供并排比较视图，迭代历史时间线可视化，快速“基于上次尝试”输入框。模式 C 界面信息丰富，任务特征面板显示复杂度、风险、熟悉度评估，情境化策略建议卡片，可靠性数据仪表板。模式 D 界面突出验证工具，一键事实检查按钮，引用验证面板，置信度指标醒目显示。模式 E 界面支持对话式教学，大文本输入框鼓励详细解释，苏格拉底式追问对话框，反思历史侧边栏。模式 F 界面包含保护性元素，确认对话框不可快速跳过，能力警告醒目置顶，独立尝试倒计时可见，使用限制通知清晰。然而界面适应是渐进和可逆的，不是剧烈重构。当模式推断改变（例如从模式 C 到模式 A），界

面元素平滑过渡，新相关工具逐渐出现，不再关键的工具淡化但保持可访问。用户始终可以手动调整界面，访问任何模式的工具，系统不强制基于模式的功能限制。适应的目标是默认提供最可能需要的工具，而不是限制访问其他工具。

元认知仪表板是第 4 层的核心组件，综合显示用户的元认知状态和发展。当前模式推断显示系统对用户当前元认知参与的理解，能力发展趋势通过图形显示随时间的独立能力变化，技能维护提醒当定期检查到期时通知，依赖警告如果检测到过度依赖模式清晰显示，个性化建议基于能力评估和使用模式提供下一步发展方向。仪表板使抽象的元认知概念具体和可操作，将“元认知发展”从模糊目标转化为可追踪进展。

解释生成根据模式和情境调整详细程度。基于第 2 层的模式推断和第 3 层的支持决策，第 4 层决定默认呈现哪个解释层次。模式 F 默认层次 1（简短摘要，他们需要简洁避免认知超载），但系统鼓励他们至少快速浏览层次 2 建立基本理解。模式 B 默认层次 1 因为在快速迭代中不需要详细解释，但在做关键方法选择时系统建议查看层次 2。模式 A 和 C 默认层次 2 提供足够细节支持战略决策和情境评估。模式 D 和 E 鼓励探索层次 3，模式 D 用于彻底验证，模式 E 用于深度学习。然而所有用户在所有情况都可以访问所有层次，系统只是智能默认。解释还适应任务风险，高风险任务自动提供更详细解释（至少层次 2）即使对通常偏好简短的模式 B 或 F，因为在高风险情况理解推理原理对所有人都至关重要。

用户控制界面确保原则 4（主体性）在实践中实现。每个 AI 建议伴随明确选项，即接受建议、修改建议（打开编辑器让用户调整）、拒绝并独立工作、请求替代方法、暂停并寻求人类专家。选项平等呈现，不通过视觉设计暗示某个选项“更好”。用户的每个选择都被尊重和支持，选择拒绝或独立工作被系统记录为积极的主体性信号，可能在能力评估中正面计分。全局设置面板让用户控制主动性水平（从“仅响应我的请求”到“主动提供建议和提醒”）、脚手架强度（从“最小支持”到“详细指导”）、透明度默认（从“简要说明”到“详细解释”）、个性化层级（影响数据收集和隐私）以及通知偏好。这些设置提供粗粒度控制，模式推断提供细粒度自动适应，两者结合实现既尊重用户偏好又智能响应实际需求的平衡。对于模式 F 用户，某些控制受到限制以保护长期能力，但限制本身是透明的并附有原理，例如“你当前无法完全禁用能力检查，因为使用模式显示可能的过度依赖，这是保护你职业能力的措施，当独立能力恢复到健康水平你将获得完全控制权”。这种有限的软家长主义是伦理上的微妙平衡，但对于主体性已经严重受损的情况是必要的。

层间交互与数据流的动态。 层间交互不是单向的数据传递而是动态的双向通信和反馈循环。向上流动始于用户在第 4 层输入任务请求，第 4 层进行初步解析提取显式信息并将请求转发到第 1 层，第 1 层执行实际 AI 计算生成响应内容同时记录过程元数据，响应内容和元数据传递到第 2 层。第 2 层分析用户行为、当前任务特征，推断或更新模式分类，评估任务复杂度、风险、熟悉度，以及判断 AI 可靠性。第 2 层的推断和评估传递到第 3 层。第 3 层基于模式和情境选择适当支持机制，实施选中的支持，并将原始响应内容、支持元素和推荐呈现方式传递到第 4 层。第 4 层综合所有元素，根据模式调整界面，应用适当的解释层次，添加用户控制元素，并最终呈现给用户。

向下流动从用户在第 4 层的行为开始。用户选择、交互模式和元认知语言都被第 4 层记录。这些行为数据传递到第 2 层作为模式分类的新证据。如果用户行为与当前模式推断一致，增强分类置信度；如果不一致，降低置信度或标记异常需要进一步观察。模式更新反馈到第 3 层调整后续支持。同时第 4 层的用户偏好告知第 3 层的个性化，即使在相同模式内不同用户可能偏好不同支持风格。用户使用第 3 层支持机制的方式提供关于支持有效性的反馈，如果用户频繁使用验证工具且发现问题表明工具有价值，第 3 层可能增加其突出度；如果用户忽略某些提示可能表明提示不相关、时机不当或表述不清，第 3 层应该调整或移除。

这种丰富的双向数据流和多重反馈循环使 MCA 系统能够持续学习和适应，不仅适应单个用户的发展（模式转变、能力增长、偏好改变），也适应用户群体的集体模式（某些任务类型普遍困难，某些支持机制普遍有效）。系统不是静态的设计产物而是动态的学习系统，通过与用户的交互不断改进其模式识别准确性、支持机制有效性和界面适应性。

3.5.4 基本设计挑战与权衡

MCA 框架在实施过程中面临五个基本设计张力。这些挑战代表固有的权衡，不能通过更好的工程完全解决，只能通过明确的设计选择和用户控制来管理。理解这些挑战及其与六种模式的关系对于实施有效的 MCA 系统至关重要。

（1）挑战 1：效率与学习深度的权衡

张力本质是即时任务完成效率与长期能力发展之间的根本冲突。最大化效率通常需要最小化认知努力，AI 快速提供答案，用户快速接受并继续下一任务，整个过程流畅无摩擦。然而深度学习需要认知努力，需要积极参与问题、与困难搏斗、犯错误并从中学习、反思过程和结果。这些学习促进活动减慢任务完成，表面上降低效率。

六种模式在这个效率-学习光谱上占据不同位置，揭示了这个权衡的复杂性。模式 B 投入大量时间进行多轮试错，不追求即时效率但追求最优解质量，他们的“效率”定义是找到最佳解决方案的效率而非最快完成任务的效率。这种质量导向的时间投入实际上促进深度学习，通过比较不同尝试用户建立对什么有效什么无效的直觉理解。模式 C 体现了动态平衡，在低风险熟悉任务中优先效率（快速接受 AI 协助，最小验证），在高风险或学习导向任务中投入深度（详细验证，理解原理）。这种灵活性表明效率和学习不是固定的个人特质而是情境化的战略选择。模式 E 明确优先学习深度，花费大量时间通过向 AI 教学来理解概念，即使这对即时任务完成是“低效”的，然而从长期看这种深度理解使未来相关任务更高效，表明效率-学习权衡有时间维度。

相反模式 F 代表极端效率导向的危险。他们最大化短期效率（快速获得答案，零反思时间）但以长期能力退化为代价。研究数据显示模式 F 用户虽然在使用 AI 时任务完成速度快，但当 AI 不可用时表现急剧下降，总体职业能力随时间衰退。他们的“效率”是虚假的，因为它依赖持续 AI 可用性并忽视能力维护成本。

MCA 应对策略不是强制所有用户到某个固定平衡点，而是支持情境化选择并保护长期利益。明确的模式识别和情境适配意味着系统识别用户当前模式并支持而非抵抗其选择。对模式 C 用户，系统提供客观的任务特征评估和成本效益分析，帮助用户做出知情的情境化决策而非替代用户判断。对模式 E 用户，系统保护其教学化反思空间，不压迫他们快速完成，如果模式 E 用户花 30 分钟向 AI 详细解释统计概念，系统不提示"你花太长时间了可以更快"而是认识到这是有价值的学习活动。

长期平衡监控通过技能退化预防（MR16）实施，对所有模式适用但对极端效率偏好者（模式 F 和某些模式 B）风险特别高。系统跟踪用户是否过度偏向效率而从不深度学习，能力是否因此停滞或下降。如果用户能力下降超过 20%，系统强制深度学习会话，这种干预可能被效率导向用户视为麻烦，但保护长期职业生存能力。值得注意的是系统承认在某些情境效率应该优先在其他情境学习应该优先，没有一刀切答案。对于重复性任务、紧迫截止日期、低学习价值活动效率优先是完全合理的，对于新领域、高复杂性、长期重要的技能学习深度应该优先即使牺牲短期速度。系统帮助用户识别哪种情境是哪种，做出明智选择，但不强加单一价值观。

（2）挑战 2：主动与被动支持的权衡

张力本质是系统发起的干预与用户发起的请求之间的平衡。主动支持意味着系统预测用户需求并主动提供建议、提醒和脚手架，优点是可以在问题发生前预防，为

不知道需要什么支持的用户提供帮助，在关键时刻提示被忽视的步骤。然而风险是侵入性，打断用户流程，可能提供不需要或不想要的支持，在极端情况下感觉像过度控制。被动支持等待用户明确请求，优点是尊重用户自主性，不干扰，仅在用户认为需要时出现。然而风险是用户可能不知道需要什么支持特别是元认知能力低的用户，问题可能在用户意识到并寻求帮助前就已经严重，依赖用户主动性而某些用户（如模式F）缺乏这种主动性。

六种模式需要显著不同的主动性水平。模式 A 需要最少主动干预，他们已经有强元认知能力系统化工作方法主动监控进度，对他们过多主动干预是干扰，受访者 I1（研究员，模式 A）评论“我欣赏系统在那里如果我需要，但我不想被不断打扰建议，我有自己的工作流程和节奏不必要的提示打断我的思考”。因此对模式 A 支持应该主要是被动的，微妙的可选提示仅在异常情况出现。模式 C 需要情境化的主动支持，在他们正确评估情境时被动支持足够，但在他们可能误判时主动"第二意见"有价值，关键是建议性语气和用户保留最终决定权。模式 E 需要主动的苏格拉底式追问来促进反思深化，但这种主动性应该被框架为学习机会而非纠正，模式 E 用户通常欢迎这种主动追问。模式 F 需要强制性主动干预不可忽略，他们缺乏元认知主动性意味着永远不会独立寻求支持，如果系统等待他们请求能力检查或依赖警告这些永远不会发生。因此对模式 F 主动性必须是强制而非可选的，虽然这违背短期便利偏好但保护长期能力。

MCA 应对策略是模式校准的主动性，即根据识别的模式动态调整主动性水平。模式 A 获得最少主动干预，模式 F 获得最多且强制的干预，模式 C 获得情境化干预，模式 E 获得学习导向的主动追问。这不是固定设置而是持续适应，如果通常模式 A 的用户开始展示有问题行为（如连续跳过验证）主动性可以暂时增加。情境敏感调整意味着主动性还取决于任务特征，高风险任务触发增加主动验证提示即使对通常最少干预的模式 A，低风险任务减少主动干预即使对通常更被动的模式 E。用户控制通过全局设置允许调整主动性偏好，但对模式 F 某些保护性干预不可禁用以保护长期利益，这个限制本身透明解释。

（3）挑战 3：通用化与个性化的权衡

张力本质是提供所有人相同体验与为每个人定制体验之间的平衡。通用化更容易实施（一个界面服务所有人，一套规则应用于所有人），需要更少数据收集（不需要了解个体特征），更好保护隐私（不收集不分析个人数据），更公平（所有人获得

相同机会，无基于数据的歧视）。然而实证发现表明用户在元认知模式上展示巨大多样性，模式 A、B、C、D、E、F 的需求定性不同，一刀切的支持不能有效服务任何人。

个性化承诺更好匹配个体需求，但需要行为数据收集（观察用户如何使用系统来推断模式），增加系统复杂性（需要模式识别算法、自适应逻辑、个性化数据库），引发隐私关注（35%参与者明确表达对数据收集和分析的担忧），以及可能的公平性问题（某些用户获得“更好”支持是否公平）。实证数据显示隐私关注的现实性和严重性，35%参与者主要是处理机密信息的专业人士明确表示在涉及敏感内容情况下不能接受任何形式的内容访问或详细行为追踪。

MCA 应对策略是分层个性化，允许用户根据情境和偏好选择适当平衡。层级 1 是最小个性化，仅使用当前会话内的行为信号（无纵向追踪），可以进行基本模式识别（从当前几次交互推断模式）和提供模式响应支持，但不跟踪长期发展或能力变化不保存个人历史。这层满足强隐私关注者需求，提供核心元认知支持（比完全通用好得多）而不需要持续数据收集。重要的是核心价值主张即模式响应性在层级 1 完全可实现，系统可以从当前交互识别用户是否展示模式 B 的迭代行为或模式 E 的教学化语言并相应调整支持，无需访问用户的完整历史或实际内容。

层级 2 是标准个性化，添加纵向模式追踪（理解用户随时间的稳定模式倾向）、能力监控（跟踪独立能力趋势）和学习路径推荐（基于进展建议下一步发展）。这需要保存用户的行为元数据（何时使用、请求什么类型任务、展示什么模式行为、能力评估结果）但仍不访问实际任务内容。这层支持脚手架淡出（需要纵向能力追踪）和能力退化预防（需要趋势监控），是大多数用户的推荐选项。层级 3 是深度个性化，添加内容感知功能如自动事实检查（需要读取 AI 输出中的具体声明来验证）、精确的引用验证和个性化学习建议。这主要惠及模式 D，对其他模式价值增量较小。这层需要明确用户同意并清楚解释收集什么数据如何使用。用户可以为不同项目选择不同层级，学术工作可能选层级 3，机密商业项目选层级 1。

隐私保护的核心功能是MCA架构的关键设计决策。最重要的元认知支持即模式识别、任务分解脚手架、元认知策略指导、能力监控和过度依赖警告都可以在无需实际内容访问的情况下运作。模式识别从行为元数据推断，脚手架和策略指导提供通用结构化支持不需要个性化到具体任务内容，能力监控使用通用练习任务和行为代理而非分析实际工作产品。这种隐私保护核心设计确保即使在层级 1 用户仍获得 MCA 框架的主要价值，不是降级体验而是功能完整但数据最小的体验。

（4）挑战 4：即时效用与长期发展的权衡

张力本质是用户为即时任务完成而使用 AI，但专门为即时效用优化可能破坏长期能力发展。用户的显式目标通常是“完成这个任务”，即时动机是速度和质量。系统如果仅响应这个即时目标会最大化 AI 协助最小化用户努力快速产生高质量输出，用户满意度在短期内很高。然而这种优化创造长期问题，技能退化因缺乏练习而积累，依赖增加使独立能力萎缩，职业价值受到威胁当用户的贡献主要是“AI 的操作员”。

相反通过拒绝即时协助或强制学习活动来优先长期发展会让用户感到沮丧并降低采用率。如果系统频繁说“在提供协助前先独立尝试 15 分钟”或“现在进行能力练习而不是完成你的任务”，用户可能放弃系统寻找更"配合"的替代品。用户生活在当下，面临真实截止日期和即时需求，长期发展虽然重要但在日常决策中往往被即时压力压倒。

六种模式在即时-长期光谱上展示不同取向。模式 E 天然长期导向，将 AI 使用视为学习机会而非仅任务完成工具，愿意投入额外时间建立深度理解，对他们长期发展本身是即时目标的一部分张力较小。模式 C 动态平衡，在特定情境专注即时完成（低风险、时间紧迫），在其他情境投入长期学习（高复杂性、战略重要性），他们的灵活性表明即时和长期不必总是冲突。模式 F 代表极端即时导向的危险，43%报告退化的用户中大多数表现模式 F 或接近 F 的行为，退化的隐匿性意味着用户没有即时反馈信号警示他们即时效用最大化策略正在长期损害他们。

MCA 应对策略是时间平衡与明确纵向监控。系统在不同时间尺度应用不同优先级。短期（单个会话内）优先任务完成提供完整 AI 协助不强制发展活动让用户高效完成手头任务，这建立用户信任和满意度。中期（每周时间框架）引入温和独立性提示通过过度依赖警告系统，提示是建议性非强制性提供意识但不强迫行为改变。长期（每月时间框架）实施强制性干预，技能退化预防每月能力检查成为不可绕过，这不是测试而是维护类似运动员保持体能。这种时间分层意味着在任何单一交互中用户体验是协助性的不感觉被限制或强迫学习，只有在累积数据显示长期趋势有问题时（能力下降、依赖增加）系统才升级到更强干预。

任务频率和重要性调整使系统智能分配长期发展压力。对于一次性罕见任务即时效用完全合理用户不需要发展永久能力，对于频繁重复的基础任务能力投资有高回报系统鼓励独立掌握，对于职业核心技能即使不常用维持能力也是关键。模式特定的能力维护反映不同模式面临不同长期风险，模式 B 用户如果总是从 AI 生成的选项中选

择可能削弱独立生成多样化方案的能力，模式 C 用户如果过度依赖系统的情境评估可能削弱独立情境判断，模式 E 用户的教学化反思本身通常保护能力但系统偶尔检查“现在独立解决这个问题（不教学给我）来确认你的理解已经内化为实践能力”。

（5）挑战 5：隐私与功能性的权衡

张力本质是许多高价值 MCA 功能需要访问用户内容，但内容访问在隐私敏感情境是交易破坏者。集成验证工具的自动事实检查需要读取 AI 输出中的具体声明来识别哪些需要验证，个性化学习路径需要理解用户实际处理什么类型任务来建议相关资源，精确的技能评估需要分析实际工作产品而非仅依赖元数据。这些功能可以显著增强用户体验，但都需要某种程度的内容访问。

然而 35%参与者明确表达隐私关注，表明内容访问在某些情境是不可协商的。受访者 I11（法律研究员）强调专业责任，律师-客户特权是法律职业的基石，不能在任何情况下允许第三方系统读取客户案件的具体细节即使是为了提供“更好服务”，这不是个人偏好而是职业道德和法律义务。受访者 I27（医生）相似地指出患者隐私受严格保护，医疗 AI 系统读取患者具体信息需要严格的法律保障不是大多数一般目的 AI 系统可以满足的，即使功能有价值隐私保护是不可妥协的。这些不是边缘关注，这些专业人士恰恰是最可能从高级 AI 协助受益的用户，如果 MCA 框架要求内容访问作为核心功能前提会排除恰恰最需要支持的用户群体。

MCA 应对策略是如挑战 3 讨论的混合架构，隐私保护核心加可选内容感知增强。核心元认知支持设计为在无需内容访问下完全运作，这不是妥协或降级而是刻意的架构选择，从一开始设计系统使隐私保护不是“事后添加的功能”而是“架构的固有属性”。模式识别从行为元数据运作，任务分解脚手架和元认知策略指导提供通用模板和清单，能力监控使用通用练习任务和行为代理，过度依赖警告从使用模式统计推断，这些都不需要分析内容。这种设计意味着即使在严格隐私要求情境用户仍获得模式响应支持、脚手架发展、能力监控和依赖警告，即 MCA 框架的四个核心原则全部可实现。

内容感知增强作为可选附加提供更精确和便利的功能但需要明确同意。自动事实检查读取 AI 输出识别声明并验证比用户手动复制粘贴到搜索引擎更快，引用验证检查来源存在性和归因准确性防御 AI“幻觉”引用问题，个性化资源推荐基于用户实际工作主题建议相关阅读、教程或专家。这些功能主要惠及模式 D 和模式 E。架构确保拒绝内容访问不会使系统无用，如果用户说“我不能允许内容访问因为处理机密信息”，系统响应“完全理解，你仍将获得完整的模式识别、任务分解脚手架、元认知策略指导、

能力监控和依赖警告，唯一的差异是自动事实检查和个性化资源推荐不可用但你可以使用集成搜索链接手动验证，核心元认知支持完全功能”。用户不会感觉被降级为“二等用户”或被惩罚隐私关注，而是做出知情的功能-隐私权衡。

项目特定隐私控制让用户为不同工作情境选择不同层级。同一律师可能在学术研究项目（无客户信息）选择层级 3 享受便利功能，在实际案件工作（机密客户数据）选择层级 1 确保绝对隐私。这种灵活性承认隐私不是个人固定特质而是情境依赖的需求，同一人在不同情境有不同隐私要求。透明数据治理支持知情选择，系统清楚解释每个层级收集什么数据、数据如何使用、数据存储多久、谁可以访问，这种透明性建立信任用户理解选择的含义。挑战的根本是技术能力（现代 AI 可以从内容提取丰富洞察）与社会价值（隐私、保密性、自主性）之间的张力，MCA 立场是社会价值优先，技术设计应该适应价值约束而非要求价值妥协以实现技术潜力，隐私保护架构虽然增加设计复杂性但使系统在道德上可接受和广泛适用，这比技术上最优但伦理上可疑的系统更有价值。

3.6 本章小结

本章通过对 49 名 AI 频繁使用者的深度访谈及系统化质性分析，完成了从实证观察到设计框架的完整转化。

在实证发现层面，本研究识别出六种彼此区分的元认知使用模式（模式 A 至 F），并揭示了一个与技术采纳领域常识相悖的核心规律：用户身份不预测 AI 使用成效。统计分析显示，教育层次与模式归属无显著关联，学科背景同样无预测力。研究生既可能表现出高度有效的战略性使用（模式 A），也可能出现问题性过度依赖（模式 F），而本科生则可能展现与资深研究者相当的元认知精细化水平。相反，元认知复杂度成为预测协作成效的关键指标，这一发现确立了有效 AI 协作本质上是元认知成就而非专业等级属性的核心论点。

在模式刻画层面，六种使用模式在元认知过程的四个维度（规划、监控、评估、调节）上呈现显著差异化特征。模式 A（战略性分解与控制）以主动任务分解与严格人类监督为核心，89%的案例显示明确的任务分解行为；模式 B（迭代优化与校准）体现容错性坚持，90%显示强调调节过程；模式 C（情境敏感的适配）展现高度的情境敏感性，同一用户在不同情境的信任标准差平均达 34.2%；模式 D（深度核验与批判性介入）实施系统性怀疑，98%显示卓越的监控行为；模式E（教学化反思与自我监控）将 AI 用作元认知发展的反思性工具，86%报告主动技能保护行为；而模式 F（被

动使用与过度依赖）则在所有维度呈现元认知缺位，虽未在本研究样本中作为主要模式出现，但根据教师二手报告，其在学生群体的实际流行率可能达 25-40%。

在设计转化层面，本章基于六种模式的实证特征，系统推导出面向元认知协作型代理（MCA）的设计框架。3.5 节建立的 MCA 概念架构以四层框架为核心：第 1 层（核心 AI 能力层）提供基础的自然语言理解、知识检索与内容生成能力；第 2 层（模式识别与推理层）负责实时检测用户的元认知模式并进行动态分类；第 3 层（元认知支持层）根据检测到的模式提供差异化支持策略；第 4 层（完整知识体系构建层）通过渐进式脚手架促进用户从低效模式向高效模式的发展性转变。这一架构不仅响应各模式用户的即时需求，更致力于实现长期的元认知能力发展。同时，3.5.4 节识别的四项基本设计挑战：效率与学习的权衡、主动干预与用户自主的平衡、模式检测的准确性要求、以及跨情境的可迁移性，为后续系统实施提供了关键考量维度。

本章建立的模式框架与 MCA 概念架构，为后续理论化讨论奠定了实证基础。第四章将通过实验验证核心设计命题，第五章将实施并评估完整的 MCA 系统。三篇论文共同构成从问题识别到设计实施的完整设计科学循环。

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Argyris, C., & Schön, D. A. (1974). *Theory in practice: Increasing professional effectiveness*. Jossey-Bass.

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology, 96*(3), 523–535. https://doi.org/10.1037/0022-0663.96.3.523

Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition: Implications for the design of computer-based scaffolds. *Instructional Science, 33*(5–6), 367–379. https://doi.org/10.1007/s11251-005-1272-9

Azevedo, R., Johnson, A., Chauncey, A., & Burkett, C. (2013). Self-regulated learning with MetaTutor: Advancing the science of learning with MetaCognitive tools. In M. Rabelo & P.

Breuleux (Eds.), *New science of learning* (pp. 225–247). Springer. https://doi.org/10.1007/978-1-4614-3657-7_11

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Article 81, 1–16. https://doi.org/10.1145/3411764.3445717

Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods & Research, 10*(2), 141-163.

Bødker, S. (1996). Applying activity theory to video analysis: How to make sense of video data in human-computer interaction. In B. A. Nardi (Ed.), *Context and consciousness: Activity theory and human-computer interaction* (pp. 147-174). MIT Press.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2022). Can explanations be good enough? How explaining AI predictions impacts trust and reliance. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1816–1826. https://doi.org/10.1145/3531146.3533248

Charmaz, K. (2014). *Constructing grounded theory* (2nd ed.). Sage Publications.

Chen, L., Chen, P., & Lin, Z. (2025). Exploring the impact of generative artificial intelligence on students' learning outcomes: A meta-analysis. *Education and Information Technologies*. Advance online publication. https://doi.org/10.1007/s10639-025-13420-z

Chilton, L. B., Kambhampati, S., Suh, J., & Teevan, J. (2024). The metacognitive demands and opportunities of generative AI. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Article 892, 1–16. https://doi.org/10.1145/3613904.3642902

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2024). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 4*, Article 100118. https://doi.org/10.1016/j.caeai.2023.100118

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Sage Publications.

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2023). How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Article 398, 1–29.

Deeva, G., De Smedt, J., De Koninck, P., & De Weerdt, J. (2021). Dropout prediction in MOOCs: A comparison study. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl

(Eds.), *Machine learning and knowledge extraction* (pp. 58–75). Springer. https://doi.org/10.1007/978-3-030-57321-8_4

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Transparency and algorithm aversion. *Management Science, 64*(1), 1–14. https://doi.org/10.1287/mnsc.2017.2955

Dourish, P. (2004). *Where the action is: The foundations of embodied interaction*. MIT Press.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. Free Press.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated rationale generation: A technique for explainable AI and its effects on human perceptions. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 263–274. https://doi.org/10.1145/3301275.3302316

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). MIT Press.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*(4), 327-358.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Flavell, J. H. (1985). Cognitive development (2nd ed.). Prentice-Hall.

Fok, R., Weld, D. S., Wu, T., Bansal, G., Kamar, E., & Ribeiro, M. T. (2024). In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine, 45*(1), 25–40. https://doi.org/10.1002/aaai.12182

Gero, K. I., & Chilton, L. B. (2019). Metaphoria: An algorithmic companion for metaphor creation. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering, 51*(12), 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research, 77*(3), 334–372. https://doi.org/10.3102/003465430303953

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18*(1), 59-82.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction, 7*(2), 174–196. https://doi.org/10.1145/353485.353487

Holtzblatt, K., Wendell, J. B., & Wood, S. (2005). *Rapid contextual design: A how-to guide to key techniques for user-centered design*. Morgan Kaufmann.

Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base. *Reuters*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Ifenthaler, D. (2024). Why explainable AI may not be enough: Predictions and mispredictions in decision making in education. *Smart Learning Environments, 11*, Article 47. https://doi.org/10.1186/s40561-024-00343-4

Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development, 68*(4), 1961–1990. https://doi.org/10.1007/s11423-020-09788-z

Kalliamvakou, E., Bird, C., Zimmermann, T., Begel, A., DeLine, R., & German, D. M. (2022). What makes a great software engineer? *IEEE Transactions on Software Engineering, 48*(3), 886–900. https://doi.org/10.1109/TSE.2020.3016771

Kalliamvakou, E., Bird, C., Zimmermann, T., Hindle, A., Abebe, S., & Nagappan, N. (2022). Programming with large language models: New programmer experiences with GitHub Copilot. *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 1497–1508.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023).

ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kim, H., Ahn, H., & Moon, J. (2021). Toward human-centered AI: A perspective on responsible AI and its societal implications. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 137–145. https://doi.org/10.1145/3461702.3462601

Kim, J., Park, S., & Lee, J. (2025). Enhancing data analysis and programming skills through structured prompt training: The impact of generative AI in engineering education. *Computers and Education: X Reality, 4*, Article 100207. https://doi.org/10.1016/j.cexr.2025.100207

Kitsantas, A., & Zimmerman, B. J. (2002). Comparing self-regulatory processes among novice, non-expert, and expert volleyball players: A microanalytic study. *Journal of Applied Sport Psychology, 14*(2), 91–105. https://doi.org/10.1080/10413200252907761

Klein, H. K., & Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly, 23*(1), 67-93.

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Leonardi, P. M. (2011). When flexible routines meet flexible technologies: Affordance, constraint, and the imbrication of human and material agencies. *MIS Quarterly, 35*(1), 147-167.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage Publications.

Lockey, S., Gillespie, N., Holm, D., & Asadi Someh, I. (2024). A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. In *Proceedings of the 54th Hawaii International Conference on System Sciences* (pp. 5463–5472). HICSS. https://doi.org/10.24251/HICSS.2021.664

Macnamara, B. N., Hambrick, D. Z., & Campitelli, G. (2024). Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cognitive Research: Principles and Implications, 9*, Article 40. https://doi.org/10.1186/s41235-024-00572-8

Malterud, K. (2001). Qualitative research: Standards, challenges, and guidelines. *The Lancet, 358*(9280), 483-488.

Maxwell, J. A. (2010). Using numbers in qualitative research. *Qualitative Inquiry, 16*(6), 475-482.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Sage Publications.

Mollick, E. R., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. *The Wharton School Research Paper*. https://dx.doi.org/10.2139/ssrn.4475995

Morse, J. M. (2015). Critical analysis of strategies for determining rigor in qualitative inquiry. *Qualitative Health Research, 25*(9), 1212-1222.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Okonkwo, C. W., & Ade-Ibijola, A. (2023). Illusion of competence and skill degradation in artificial intelligence-driven programming education. *International Journal of Research in Science and Innovation, 10*(5), 1725–1738. https://doi.org/10.51244/IJRSI.2023.1005134

Orlikowski, W. J. (2000). Using technology and constituting structures: A practice lens for studying technology in organizations. *Organization Science, 11*(4), 404-428.

Park, J., Kim, S., & Lee, H. (2025). The impact of individual AI proficiency on human-agent collaboration: Higher sensitivity to discern the comprehension ability of intelligent agents for users with higher AI proficiency levels. *International Journal of Industrial Ergonomics, 105*, Article 103514. https://doi.org/10.1016/j.ergon.2025.103514

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Sage Publications.

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv preprint arXiv:2302.06590*. https://arxiv.org/abs/2302.06590

Pew Research Center. (2024). *Americans' use of ChatGPT is ticking up, but few trust its election information*. https://www.pewresearch.org/short-reads/2024/02/26/americans-use-of-chatgpt-is-ticking-up-but-few-trust-its-election-information/

Prunkl, C., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence, 3*(2), 104–110.

Puentedura, R. R. (2006). Transformation, technology, and education [Blog post]. http://hippasus.com/resources/tte/

Rahiman, H. U., & Kodikal, R. (2023). Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications, 10*, Article 311. https://doi.org/10.1057/s41599-023-01787-8

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Rossi, P. G., & Ferri, P. (2025). Mastering knowledge: The impact of generative AI on student learning outcomes. *Studies in Higher Education, 50*(1), 1–14. https://doi.org/10.1080/03075079.2025.2487570

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*(1), 342–363.

Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). Sage Publications.

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2024). Psychological traits and appropriate reliance: Factors shaping trust in AI. *International Journal of Human-Computer Interaction, 41*(7), 1657–1673. https://doi.org/10.1080/10447318.2024.2348216

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*(4), 460–475. https://doi.org/10.1006/ceps.1994.1033

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*(4), 351–371. https://doi.org/10.1007/BF02212307

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management, 31*(2), 74–87. https://doi.org/10.4018/JDM.2020040105

Strzelecki, A. (2023). What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence, 5*, Article 100449. https://doi.org/10.1016/j.caeai.2023.100449

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning and Teaching, 6*(1), 31–40.

UNESCO. (2023). *ChatGPT and artificial intelligence in higher education: Quick start guide*. UNESCO. https://www.iesalc.unesco.org/en/2023/04/14/chatgpt-and-artificial-intelligence-in-higher-education-quick-start-guide/

Vaccaro, K., Almaatouq, A., & Malone, T. W. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour, 8*(11), 2109–2123. https://doi.org/10.1038/s41562-024-02024-1

Van der Stel, M., & Veenman, M. V. J. (2008). Relation between intellectual ability and metacognitive skillfulness as predictors of learning performance of young students performing tasks in different domains. *Learning and Individual Differences, 18*(1), 128–134. https://doi.org/10.1016/j.lindif.2007.08.003

Van der Stel, M., & Veenman, M. V. J. (2010). Development of metacognitive skillfulness: A longitudinal study. *Learning and Individual Differences, 20*(3), 220–224. https://doi.org/10.1016/j.lindif.2009.11.005

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), Article 129, 1–38. https://doi.org/10.1145/3579605

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*(1), 3–14. https://doi.org/10.1007/s11409-006-6893-0

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information Systems Research, 3*(1), 36-59.

Walsham, G. (1995). Interpretive case studies in IS research: Nature and method. *European Journal of Information Systems, 4*(2), 74-81.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1990). What influences learning? A content analysis of review literature. *Journal of Educational Research, 84*(1), 30–43. https://doi.org/10.1080/00220671.1990.10885988

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). Springer-Verlag.

Weick, K. E. (1995). *Sensemaking in organizations*. Sage Publications.

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences, 8*(4), 327–353. https://doi.org/10.1016/S1041-6080(96)90022-9

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Lawrence Erlbaum Associates.

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, 75–78. https://doi.org/10.1145/3581754.3584136

Xu, W., Jang, Y., & Ouyang, F. (2024). AI-empowered self-regulated learning: A systematic review. *Computers and Education: Artificial Intelligence, 6*, Article 100247. https://doi.org/10.1016/j.caeai.2024.100247

Young, A., & Fry, J. D. (2008). Metacognitive awareness and academic achievement in college students. *Journal of the Scholarship of Teaching and Learning, 8*(2), 1–10. https://scholarworks.iu.edu/journals/index.php/josotl/article/view/1696

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2024). Rethinking trust calibration in human-AI collaboration. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Article 345, 1–18. https://doi.org/10.1145/3613904.3642157

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2