MentalGame：利用多类型游戏和机器学习方法预测软件开发人员的人格与工作适配度

Soroush Elyasi[a]、Arya VarastehNezhad[a] 和 Fattaneh Taghiyareh* [a]

[a]德黑兰大学计算机工程系

arXiv:2601.01206v1 [cs.LG] 3 Jan 2026

摘要

职业指导和人员选拔中的人格评估，传统上依赖自我报告问卷，但这类问卷容易受到反应偏差、疲劳和有意歪曲的影响。基于游戏的评估通过捕捉游戏过程中的隐式行为信号，提供了一种前景广阔的替代方案。本研究提出一种多类型严肃游戏框架，并结合机器学习技术，以预测候选人对软件开发岗位的适合度。我们通过系统性强的文献综述和对专业软件工程师的实证研究，确定了与开发人员相关的人格和行为特质。我们设计了一款定制手机游戏，用以引发与问题解决、规划、适应能力、坚持性、时间管理和信息搜寻相关的行为。研究收集了细粒度的游戏事件数据，并采用两阶段建模策略进行分析，仅依据源自游戏过程的行为特征预测适合度。结果显示，我们的模型最高达到97%精确率和94%准确率。行为分析表明，合适的候选人呈现出独特的游戏玩法模式，例如在益智类游戏中获胜更多、参与更多支线挑战、更频繁地浏览菜单，以及更少出现暂停、重试和放弃行为。这些发现表明，游戏过程中捕捉到的隐式行为轨迹有望在不进行显式人格测试的情况下预测软件开发岗位适合度，也说明严肃游戏可作为一种可扩展、更具吸引力且偏差更少的职业评估替代方案。

关键词：基于游戏的评估、严肃游戏、行为分析、职业指导、人机交互、迈尔斯-布里格斯类型指标（MBTI）

1. 引言

人格评估在当代各种决策过程中发挥着至关重要的作用。基于人格数据得出的洞察，常用于信用风险评估、员工招聘、工作满意度和任务绩效预测、团队构成以及教育指导。[1–6] 在许多此类情境中，研究人员会分析各种形式的行为数据，从社交媒体轨迹到数字活动日志，以发现有助于构建可靠预测模型的模式。[7,8] 然而，尽管人格问卷被广泛用作主要评估工具，这些工具仍存在若干结构性局限。自我报告偏差、疲劳、对题目的误解以及刻意的印象管理，都会削弱基于问卷的评估的准确率和公平性。[9–14] 因此，寻找能够捕捉真实行为倾向和潜在心理范式的替代方法，已成为一项重要的研究目标，并促使研究人员探索新的途径。[15–17]

电脑游戏是此类替代性评估最有前景的途径之一。[1,12,18] 游戏趣味十足、引人入胜，而且通常不会带来正式测试环境所伴随的心理压力。[19] 玩家会自发地进行互动，做出不受约束的决策，并展现出行为模式；这些模式往往能更直接地反映他们真实的偏好、能力和倾向。[18] 游戏能够记录细致的事件日志、移动模式、决策点、反应时间、探索选择、不必要或可选的操作以及犹豫，这使其能够提供极为丰富的行为数据集。[14, 20–23] 由于这些行为痕迹是在潜意识状态下产生的，基于游戏的评估能够减少传统自我报告形式固有的许多问题。这促使我们深入研究游戏作为人格和行为分析可靠媒介的潜力。

此前的大量研究考察了玩家在《上古卷轴V：天际》、《战地3》、《英雄联盟》、魔兽世界、《我的世界》和《Dota 2》等广为人知的商业游戏中的决策与人格特质之间的联系。[18, 24–26] 其他研究则设计了简单的游戏，通常以文字或益智类游戏为基础，并将由人格驱动的机制直接融入游戏过程。总体而言，这些研究有力地表明，游戏行为与多种心理构念相关，包括合作倾向、冲动性、问题解决策略、情绪稳定性，甚至游戏成瘾的易感性。[25, 27–31]

尽管取得了这些进展，仍存在一些重大不足。首先，现有研究大多以五因素模型（大五人格）作为基础心理框架，而迈尔斯-布里格斯类型指标（MBTI）虽然广泛应用于组织和职业安置情境，却受到的关注少得多。[9,32,33]2018—2023年间，以大五人格为基础的研究占据主导地位，而基于MBTI的研究相对稀少（表1），这表明相关文献存在明显失衡。其次，尽管已有研究探讨将游戏用作人格推断工具，但很少有研究利用游戏数据开展实际的职位推荐或人员选拔任务，更少有研究关注特定职业。针对现实应用的研究，例如识别适合特定岗位的候选人，无论在研究范围还是方法多样性方面都仍然有限。第三，现有研究很少将人格问卷与某一特定职业可能涉及的其他行为特质结合起来。这为设计更全面、更具领域针对性的评估框架提供了重要契机。总体而言，本研究的贡献主要体现在三个方面。

1. 本文全面综述了基于游戏的人格评估文献，重点介绍了自2010年以来的方法论模式、问卷选择以及研究趋势的演变。 2. 本文提出了一种新颖的基于游戏的评估框架，将迈尔斯-布里格斯类型指标与专门针对软件开发人员的互补行为特征相结合，弥补了现有研究中的一个空白。 3. 本文利用游戏事件数据开发了一个可操作的预测模型，证明了在游戏过程中捕捉到的行为轨迹可用于识别具有软件开发人员潜在能力倾向的个体。

开发岗位。

图1展示了本研究的图形摘要。

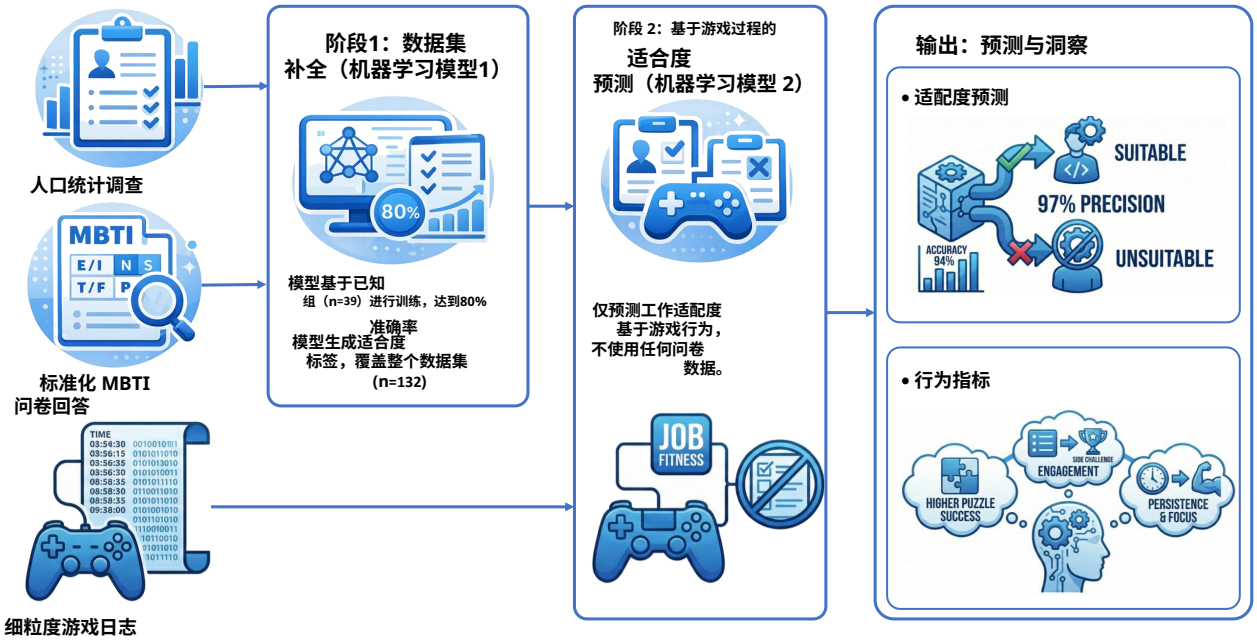

图1. 研究的图形摘要

2. 相关工作

心理测验长期以来一直被用作评估个体行为类型和认知能力的科学工具。随着时间的推移，这些测验已实现标准化，通常通过问卷进行。 然而，基于问卷的评估存在若干众所周知的局限，包括受访者疲劳、反应偏差、焦虑、有意或无意的错误回答，以及社会赞许性反应。[34]尽管存在这些缺点，此类评估仍受到雇主的青睐，被用于人员选拔和团队组建。[35] 它们还广泛用于诊断心理障碍，以及设计教育与心理改善项目。[12,36]

基于问卷的评估面临一项根本挑战：它依赖自我报告，而这不可避免地会带来偏差。受访者可能误解或看错问题，而且问卷往往被认为枯燥乏味。实证研究表明，当人们参与自己觉得乏味的任务时，注意力会下降，进而给出随意或不切实际的回答。[12,34] 这些缺点促使人们探索其他评估方法。

游戏化和基于游戏的评估方法旨在将评估机制嵌入游戏过程，以缓解传统问卷的局限性。[12] 通过在游戏过程中隐式收集行为数据，这些方法能够减少偏差、提高参与度，并得出更可靠的结果。[12,34]

大量研究已证实，游戏内行为、玩家选择、表现指标与人格特质之间存在有意义的关系。[37,38] 针对《上古卷轴V：天际》、《战地3》、《英雄联盟》、魔兽世界、《我的世界》和《Dota 2》等热门游戏的研究，证实了人格特质与游戏内行为之间存在显著关联。[37] 其他研究则侧重于设计与心理学框架相契合的定制游戏，尤其是基于五因素模型（大五人格）的游戏，并成功发现了行为特质、问题解决策略与游戏玩法模式之间的显著相关性。[22,23,39]

Zulkifly开发了一款游戏，旨在分析玩家的问题解决方法、地图探索行为、暂停时长、移动模式（必要和非必要），并根据这些行为推断人格特质。[22] 另一项研究采用了一款文本游戏，其中每个问题都提供三个回答等级（低、中、高）。该方法旨在通过文本式游戏玩法对大五人格特质建模，与传统方法相比，其结果与问卷结果的吻合度更高。[12] Haizel等人推出了一款角色扮演游戏，其中包含一个简短的奇幻主题叙事，并嵌入了12个决策点。尽管可选项数量相对较少，但平均而言，这款游戏对人格的估计比传统的游戏后问卷更准确。[23]

另一项研究结合NEO-PI问卷与一款定制游戏，邀请44名玩家参与，并使用275个变量识别五种人格特质。结果表明，基于游戏的输出与基于问卷的评估之间存在很强的相关性。[39] 另一款基于Unity的游戏在30名玩家中进行了测试，且未向参与者透露评估目的。该游戏的持续时间长于传统问卷，并使用NEO-PI进行验证。结果显示，玩家行为与问卷结果之间存在直接关系，而且与传统评估相比，参与者满意度更高。[40]

Gao等人研究了现有在线游戏中的人格识别。在他们的首项研究中，他们对《列王的纷争》进行了分析，采用K均值聚类，随后进行统计相关性分析。研究使用了HEXACO人格框架，得出以下发现：[37]

- 外向性较高的玩家表现出更多攻击性行为。
- 尽责性与游戏内死亡呈负相关。
- 情绪性呈现出显著的行为效应：高分或低分均与杀戮和攻击行为相关。
- 诚实-谦逊、宜人性和开放性无法显著预测游戏内行为。

另一项研究聚焦于射击游戏，采用五因素模型和机器学习技术来预测人格特质。研究人员开发了一款名为 Protector 的定制游戏，其中人工神经网络取得了最高准确率，而支持向量回归和K近邻回归的预测误差最低。[38] 多项研究比较了视觉游戏与文字类游戏在人格评估方面的效果，发现两者效果相当；与问卷相比，玩家对基于游戏的方法反应更积极。[41] 其他研究强调了游戏在减少偏差和尽量避免错误作答方面的作用，尤其是与大五人格框架相契合的基于决策的冒险游戏。[42]

进一步的研究探讨了面向选拔的游戏，并得出结论：设计精良的人格评估游戏可有效用于高风险选拔情境。该研究采用了 HEXACO 框架，并重复使用了此前开发的一款游戏。[35] 彼得大帝大学开展的一项为期五年的研究，运用站点、测验和 Instagram竞赛等游戏化元素，分析一年级学生的人格表现。[43]

研究者还提出了密室逃脱类游戏，利用大五人格模型，通过多样化的情境和数学建模提取人格特质。[44] 其他研究采用逻辑回归和决策树模型，协助雇主进行候选人选拔；研究发现，通过游戏提取认知预测比提取人格特质更准确。[45] 研究者还引入了虚拟现实环境，以模拟真实的社会互动，进行人格评估。[46] 另有研究考察了竞技游戏中的人格与角色选择之间的关系，包括性别效应。研究使用了 208 名成年玩家的数据，结果显示，没有证据可以拒绝关于辅助角色玩家宜人性更高的原假设。[47]

人们对游戏化评估日益关注，原因在于它能够减少作弊行为，并引发候选人的积极反应。理论观点认为，游戏化可能通过增加认知负荷并掩盖所测量的特质来抑制作弊，从而使候选人更难故意以过度有利的方式展现自己。一项研究试图探究这一点，将传统人格评估与游戏化人格评估结合起来，并明确警告参与者，不诚实回答是可被检测出来的，以考察隐性游戏化测试与回答诚实性之间的关系。[48] 其他小规模研究则利用大五人格模型，证实了游戏行为与人格特质之间存在稳定的相关性。[49]

表1总结了2019年至2023年间发表的相关研究，包括所采用的人格问卷、游戏类型和分析方法。对于未明确说明分析方法或游戏类型的研究，不纳入百分比计算。正如图2所示，所综述的研究大多采用大五人格框架。此外，该图表明，大多数研究更倾向于使用自行开发的游戏，而非商业游戏；在评估方法中，统计分析占主导地位，其次是问卷到游戏映射和机器学习方法。

3. 方法论

3.1 与开发者相关的特质识别

所提出方法论的第一阶段重点在于识别与软件开发联系最为密切的人格特征和行为特质。这一阶段至关重要，因为它为游戏设计提供了概念基础，并决定了分析过程中哪些行为被视为具有意义。为此，本研究采用双重分析策略，将现有国际研究的系统综述与在伊朗开展的独立实证研究相结合。

对既有研究的综述显示，关于软件工程师人格的研究往往依赖学生样本，而非在职专业人士。[5,10,34,56] 尽管此类研究提供了有价值的洞察，但可能无法准确反映实际从事软件开发岗位人员的特征。此外，文化和社会因素可能会显著影响人格分布和行为表现。[57,58] 基于这些原因，本研究对受雇于伊朗公司的专业软件工程师开展了实证研究，旨在验证文献中报告的研究结果，并结合具体情境对其进行解读。[10]

对国际研究和伊朗数据集的分析表明，将人格框架视为僵化的分类系统（例如完整的迈尔斯-布里格斯人格类型），解释力有限，

表1. 基于游戏的人格评估研究（2019—2023年）

Ref.	Year	人格模型	游戏来源	分析方法
[18]	2023	MBTI	自制	关系（统计分析）
[37]	2023	HEXACO	商业	机器学习
[38]	2023	大五人格	自制	机器学习
[41]	2022	大五人格	自制	问卷条目到游戏机制
[42]	2022	大五人格	自制	关系（统计分析）
[35]	2022	HEXACO	自制	关系（统计分析）
[43]	2022	MBTI	自制	关系（统计分析）
[50]	2022	大五人格	自制	关系（统计分析）
[1]	2022	None	None	None
[44]	2022	大五人格	自制	问卷条目到游戏机制
[45]	2021	大五人格	自制	机器学习
[40]	2021	大五人格	自制	关系（统计分析）
[46]	2021	None	自制	None
[23]	2021	大五人格	自制	问卷条目到游戏机制
[47]	2021	大五人格	自制	关系（统计分析）
[48]	2021	HEXACO	自制	关系（统计分析）
[3]	2021	迈尔斯-布里格斯类型指标 / 大五人格	None	None
[49]	2020	大五人格	自制	关系（统计分析）
[51]	2020	HOSP	自制	关系（统计分析）
[52]	2020	大五人格	自制	问卷条目到游戏机制
[53]	2020	多项（DSM-5）	None	关系（统计分析）
[2]	2020	None	None	None
[54]	2019	大五人格	商业	关系（统计分析）
[55]	2019	未提及	商业	关系（统计分析）
[12]	2019	大五人格	自制	关系（统计分析）
[22]	2019	大五人格	自制	问卷条目到游戏机制

尤其是在样本量有限的情况下。[10,34] 相反，对人格的独立维度进行分析，能够提供更稳健且在统计学上更有意义的表征。[10,56,57] 具体而言，基于迈尔斯-布里格斯类型指标得出的结果始终表明，思维与情感维度与软件开发岗位的关联最强、最稳定。[5,10,34,56] 研究发现，偏好分析性、逻辑驱动的决策者在开发人员中占比显著偏高，而主要依据情感或人际因素做出决策的人占比偏低。思维型（T）个体更为突出，也更适合从事软件开发，因此应当加以识别。这类个体具有以下特征：[59,60]

- 逻辑性强
- 分析性强
- 判断时不掺杂个人情感（优先考虑真相而非圆融）
- 任务导向（基于规则）且系统性强

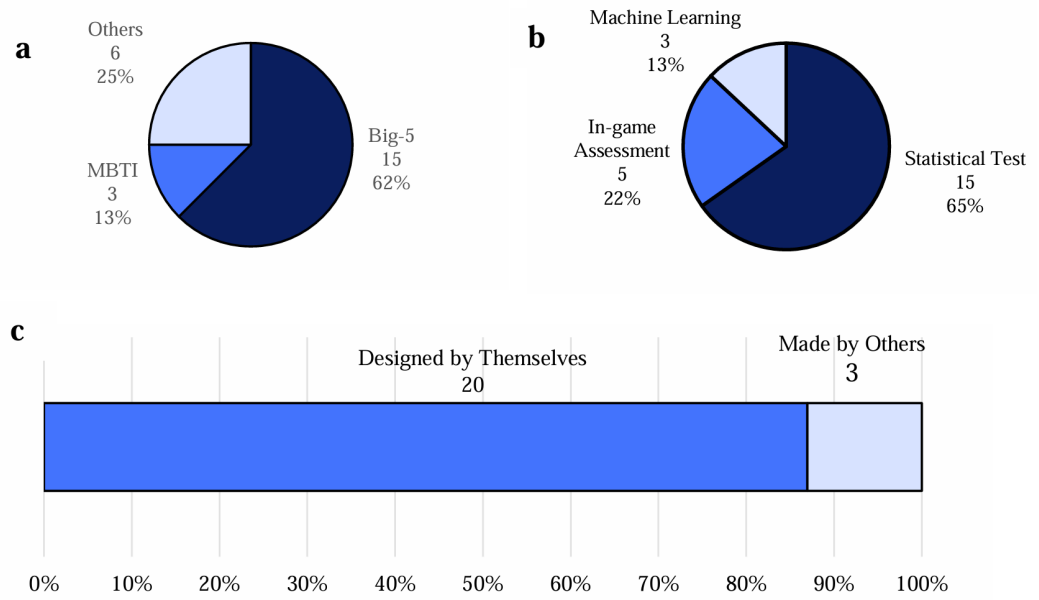

图 2. 纳入综述的既有基于游戏的人格评估研究概览（2019—2023年）。（a）所用的人格模型。（b）采用的分析方法。（c）研究所用游戏的来源。

- 以问题解决为导向
- 批判性与怀疑性思维
- 构思、创造力与创新
- 数据驱动

除了这一主要人格维度之外，文献和实证数据中还反复出现与软件开发相关的行为特质。为确保稳健性，我们优先考虑在所审阅资料中出现次数超过三次的特质。为减少冗余，我们合并了概念上相互重叠的特质，尤其是那些经常与迈尔斯-布里格斯类型指标思维型特征相关联的特质。因此，我们从基于文献得出的特征最终列表中剔除了冗余特质，形成了一组更连贯且便于实际测量的特质，包括系统性问题解决、逻辑推理、对结构和细节的关注、适应新挑战、有效的时间管理、持续学习，以及在必要时寻求信息或帮助的能力。[57,58,61_64]成就动机以及乐于应对复杂或有挑战性任务的意愿也经常被提及。[65–67]总体而言，这些特质为潜在行为构念的选择提供了依据；严肃游戏旨在引发并测量这些构念。

总体而言，结果表明，适合的软件开发人员主要表现为 MBTI 框架中的 T 特质，并辅以关键的行为属性。具体而言，寻求信息和帮助、有效管理时间、保持成功动机、适应不断变化的条件以及参与持续学习的能力，对于在软件开发中实现有效且可持续的绩效至关重要。

3.2 技术基础

方法论的下一阶段是设计并实现一款严肃游戏，将其作为主要的行为数据收集工具。该游戏使用 Unity 游戏引擎开发，该引擎因其灵活性、跨平台能力以及对仪器化研究应用的适合度而被选用。所有游戏过程逻辑和交互处理均使用 C# 编程语言实现。

选择 Android 平台进行部署，以最大限度地提高可访问性，并确保不同背景的参与者无需专用硬件即可轻松安装和使用该应用程序。为支持详细的行为日志记录及后续分析，我们开发了专用的后端基础设施。

该基础设施采用 ASP.NET Core 实现为基于 Web 的应用程序编程接口，负责接收来自客户端设备的游戏过程事件。Microsoft SQL Server 被用作主要的数据存储系统，为高容量事件数据提供可靠且结构化的存储。我们使用 Entity Framework 定义数据模型，并管理应用程序与数据库之间的交互。

参与者执行的每项游戏过程操作，包括移动、选择、暂停、重试和导航事件，均被记录为结构化事件并传输至后端服务器。数据传输过程受到安全保护，且所有参与者都在参与前获知所收集数据的性质。在进行任何数据收集之前，均已取得参与者的明确同意。游戏数据集中不存储任何个人身份信息，从而确保参与者隐私并符合伦理研究标准。

3.3 设计策略与游戏结构

首要目标并非单纯提供娱乐，而是设计出需要玩家付出认知努力、运用战略思维、适应变化并在限制条件下学习的游戏情境。最初，我们探索了两个主要的游戏概念。第一个是受高决策性机制启发的叙事驱动的卡牌角色扮演游戏，玩家需要做出影响故事推进的选择。尽管这一设计提供了丰富的定性数据，但试点测试表明，玩家对叙事的投入和情感参与往往会盖过分析性决策。玩家行为深受其对故事结局的个人偏好影响，而非问题解决策略，这降低了根据记录的操作来解读软件开发相关特质的能力。因此，这一原型未纳入最终设计。

第二个原型是一款塔防类游戏，着重考验策略性布置、资源管理和长期规划。尽管这一设计展现出测量规划能力的潜力，却带来了相当陡峭的学习曲线。许多参与者在游戏过程中花费了大量时间理解游戏机制，而非投入有意义的问题解决。考虑到实验时间有限，且需要尽量减少学习偏差，这个原型也被舍弃了。

基于这些观察，最终的设计策略采用了模块化多阶段方法，由简短且彼此独立的游戏场景组成。每个阶段都经过设计，便于快速上手，同时随着难度提升，仍要求参与者运用愈加复杂的推理能力。这种方法使研究者能够在单次游戏过程中评估多种认知与行为特征，同时尽量减少疲劳、挫败感和混杂学习效应，以及参与者对类似游戏机制的先验知识所带来的影响。尤为重要的是，各阶段彼此独立，确保参与者即使在某项任务中失败或遇到困难，也不会因此无法继续参与后续任务。

最终定稿的游戏由多个各具特色的阶段组成，每个阶段都旨在针对与软件开发相关的特定行为构念。这些阶段涵盖多种类型，包括受限解谜、基于记忆的推理、压力下的实时行动、空间操控和抽象图遍历。尽管各阶段的游戏机制各不相同，但都遵循一致的设计理念：每项任务都会施加显式约束，促使玩家进行推理、规划、检验假设，并根据反馈调整策略。

其中一类阶段侧重于受限问题解决任务：玩家必须在有限空间内重新排列物体或实体，以达到预定义目标状态。这些任务旨在衡量系统性规划、前瞻性，以及在执行前推理行动序列的能力。随着难度增加，解空间呈组合式扩张，使蛮力方法变得效率低下，并促使玩家运用分析性推理。玩家在这些阶段的行为有助于了解他们如何应对复杂问题：是提前规划，还是依赖试错；以及当初始策略失败时，他们会如何应对。

另一类关卡以基于记忆的挑战为核心。在这些关卡中，玩家会短暂接触一些信息，之后必须在时间压力下回忆或推断这些信息。虽然短期记忆容量会发挥作用，但成功表现越来越取决于策略的形成，例如模式识别或系统性探索。设置这些关卡并非为了单独评估记忆力，而是为了观察参与者面对不完整信息时如何调整自己的方法；不完整信息是现实世界软件开发任务的常见特征。

另一组关卡涉及实时生存场景，玩家必须持续应对动态变化的状况。这些关卡要求玩家迅速决策、安排行动优先级，并适应突发事件。与谜题类关卡不同，这些场景提供了更大的移动自由度和选择空间，因此能够观察个体在风险容忍度、情境意识以及从反复失败中学习方面的差异。这些关卡所产生的丰富行为数据，有助于提取多种与适应能力及压力下决策相关的特征。

其他关卡侧重于空间推理和抽象思维。在空间操控任务中，玩家必须重新摆放障碍物或元素，引导实体到达目标；这需要预测未来状态并评估备选配置。在图遍历任务中，玩家面对的是受特定遍历规则约束的互连结构，必须找到满足全局约束的路径。这些关卡旨在引发更高层次的推理、心理模拟以及管理复杂规则集的能力，而这些能力都是算法设计和调试等软件开发活动的核心。

在所有阶段中，我们有意采用了若干贯穿整体的设计机制，以捕捉有意义的行为信号。我们设置了时间限制，以模拟现实约束并观察参与者如何应对压力。每个阶段开始时都提供可选教程，让参与者可以选择指导式教学或探索式学习。这一选择本身就能体现其学习偏好和信心。此外，我们还加入了暂停和恢复游戏过程的功能，以捕捉策略性退出、反思和规划行为。

一项尤为重要的设计元素是加入了自愿放弃选项。当玩家超过某个阶段的规定时间时，可以选择继续尝试任务，也可以放弃并进入下一阶段。这一机制可以衡量玩家在感知失败时的坚持性、挫折耐受力和决策能力。该方法并未将放弃视为纯粹的负面结果，而是将其视为一种有意义的行为信号；具体而言，放弃可能反映适应性的自我调节，也可能表明回避，具体取决于情境和发生频率。

为进一步丰富行为数据集，游戏中加入了可选的支线挑战。这些挑战通常基于数学或逻辑问题，玩家完成后可以获得资源，用于跳过特别困难的阶段。参与这些可选任务有助于了解玩家的内在动机、投入额外努力的意愿，以及对学习和自我提升的态度。重要的是，玩家无需参与这些挑战，从而确保参与体现的是自愿行为，而非服从。

游戏的整体结构旨在兼顾控制与自由。每个阶段都有特定的规则和限制，但玩家仍可自主选择解决问题的方式。这种平衡确保收集到的数据能够反映玩家在认知风格和行为策略上的真实个体差异，而非僵化机制所导致的千篇一律的反应。有关游戏及其条件的更多详情，请参见附录A。

3.4 实验程序与数据收集

游戏设计和系统实现完成后，方法论的下一阶段便是执行实验程序并系统性地收集数据。此阶段旨在确保收集到的行为数据和心理数据可靠，且适用于后续的统计分析和机器学习分析。实验过程分为两个主要阶段：试点研究和最终的大规模实验，各自承担不同的方法论目的。

这项试点研究作为初步验证步骤开展，旨在评估游戏的可用性、说明的清晰度以及事件记录机制的信度。这项初步实验共有34名参与者，他们来自不同群体，包括专业软件工程师、计算机相关学科的研究生，以及没有计算背景的人士。纳入背景各异的参与者，有助于观察多种交互模式，并发现可能对某些用户群体造成不成比例影响的潜在设计缺陷。游戏平台

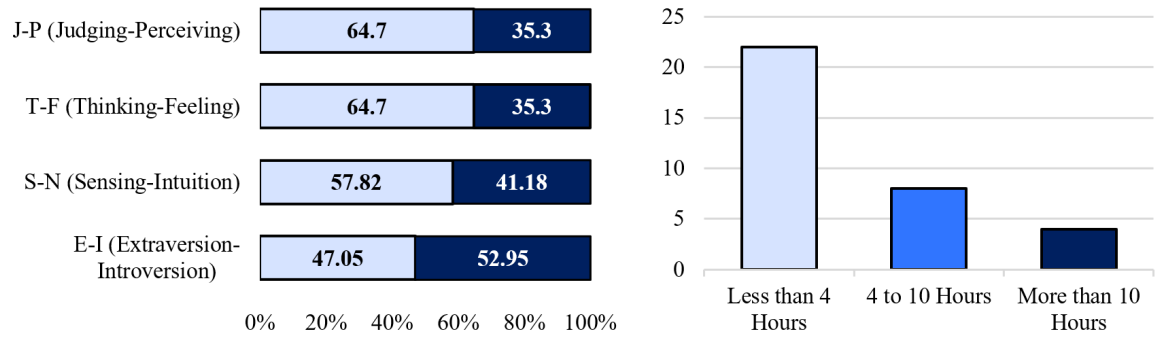

图3. 迈尔斯-布里格斯类型指标各字母的分布，以及玩游戏的平均时长

偏好、平均游戏时长，以及第一项实验中参与者的迈尔斯-布里格斯类型指标字母和配对类型分布，见图3。

在试点研究中，参与者完成了完整的游戏流程和相应的人格问卷。观察结果和非正式反馈表明，虽然游戏的整体结构易于理解，但某些阶段需要更清晰的说明，以减少初期的困惑。根据这些发现，我们完善了教程内容，强化了视觉提示，并调整了难度递进，以确保早期阶段的挑战成为学习机会，而非障碍。此外，我们改进了行为日志记录的细致程度，以确保能够一致地捕捉细微动作，例如移动前的犹豫或反复探索行为。试点研究还提供了初步证据，表明该游戏能够引发参与者之间有意义的行为差异，从而支持其用于主要实验的适合度。

根据试点阶段的结果进行改进后，我们开展了最终实验研究。本阶段共招募了153名参与者。参与者来自中学、大学和职业环境，确保了其年龄、教育背景以及对计算概念的熟悉程度具有多样性。为确保人格评估工具的效度，参与者必须年满14岁。我们获得了伦理审批，并在数据收集前取得了所有参与者的知情同意。此外，在游戏的最后阶段，他们可以选择将匿名化数据发送至我们的服务器，或选择不提交。如果选择发送数据，他们会收到一段代码，并在提交问卷时一并提供，以便将两者关联起来。每位参与者都完成了实验包含的三个部分。

- 他们填写了标准化的迈尔斯-布里格斯人格问卷，该问卷提供了独立的人格维度得分，用于后续分析。参与者被分为两组：一组在进行游戏前填写问卷，另一组在游戏结束后填写问卷，以减少两者之间潜在的相互影响。

- 参与者完成了一项人口统计与自我评估调查，内容涵盖教育程度、职业经历、对软件开发的兴趣，以及对自身完成技术任务能力的认知等信息。
- 参与者完成了全部游戏阶段，在此过程中，后端系统记录了所有交互。

数据收集完成后，对数据集的完整性和质量进行了检查。包含不完整问卷回答、中断的游戏环节或损坏事件日志的记录均被排除在进一步分析之外。经过此次清理后，剩余132条参与者记录，并被认定为后续建模的有效记录。该最终数据集构成了方法论其余阶段的实证基础。

游戏环节生成了一个高维行为数据集，捕捉参与者互动的细微特征。最初提取了 132 个不同的源自游戏过程的变量。这些变量涵盖了广泛的行为，包括但不限于操作频率、

重试次数、错误数量、每个关卡所花的时间、教程的使用情况、暂停频率、放弃次数、在菜单中的导航情况，以及参与可选支线挑战的情况。每个变量的设计都旨在直接或间接地对应本研究前文确定的一项或多项与开发者相关的特质。这些变量的列表见附录 B。

为准备数据集以供分析，我们进行了全面的预处理。为避免在后续模型中引入噪声和不稳定性，我们移除了数据覆盖率不足的变量，例如对应于极少触发事件的变量。我们也排除了在参与者之间近零方差的特征，因为这类特征几乎不提供判别信息。此外，我们进行了相关性分析，以识别可能造成冗余并对模型性能产生负面影响的高度相关变量。若观察到较强的相关性，则保留可解释性更强或理论上更相关的变量。完成这些预处理步骤后，最终保留了约一百二十个行为变量。

在预处理行为数据的同时，我们实施了一套标注策略，以支持监督学习。我们根据部分参与者在软件开发领域已得到确认的专业经验对其进行标注，包括经核实的学术或职业参与经历。核实工作结合了自我报告的问卷回答与专家判断。约四十名参与者符合这些标准，并被赋予二元标签，以表明其适合或不适合从事软件开发。这些已标记实例作为模型训练和评估的真实标注。其余参与者则被视为未标注样本，由预测模型推断其适合度。

这种标注方法反映了现实职业指导场景中的实际限制：明确的标签往往稀缺且获取成本高。该方法依赖数量有限但经过精心筛选的已标注样本，旨在兼顾现实性与分析严谨性。未标注数据的存在也促使我们采用稳健的建模技术，以便在监督信息有限的情况下实现良好的泛化。

这一阶段结束时，数据集由三个紧密整合的部分组成：根据 MBTI 问卷得出的独立人格维度、人口统计变量和自我评估变量，以及从游戏过程中提取的高维行为表征。这些部分共同构成了丰富而多元的基础，为下一节所述的机器学习分析提供了支持。第二次实验中玩家的人口统计信息和 MBTI 人格档案见图 4。

3.5 机器学习框架与建模策略

完成数据收集、预处理和标注后，研究方法进入分析与建模阶段。该阶段旨在构建预测模型，根据收集到的人格数据（重点关注 T 型人格特质）、其他问卷回答，以及一小部分经核实的适合与不适合案例，推断个体对软件开发的适合度。其他问题依据文献中常与 MBTI 框架下 T 型特质相关联的特征设计；但为避免重复评估，明确排除了与 T 维度本身直接重叠的特征，因为 T 人格维度已用作自变量。

为以结构化且可解释的方式实现这一目标，该机器学习框架被划分为两个连续阶段。第一阶段侧重于利用基于人格的信息完善带标签数据集，第二阶段则专注于仅使用从游戏玩法中提取的行为特征，在整个数据集上重新预测适合度。

第一个建模阶段旨在解决数据集补全问题。鉴于只有部分参与者具有明确的标签，用以表明其是否适合软件开发，本阶段旨在利用心理学信息和自我报告信息，预测其余参与者的标签。该任务被设定为二元分类问题，每位参与者都会被划分为适合或不适合两类之一。本阶段的输入特征主要由从迈尔斯-布里格斯问卷中提取的独立人格维度构成，并辅以少量自我评估变量，用于反映学习行为、适应能力、时间管理和信息寻求倾向。

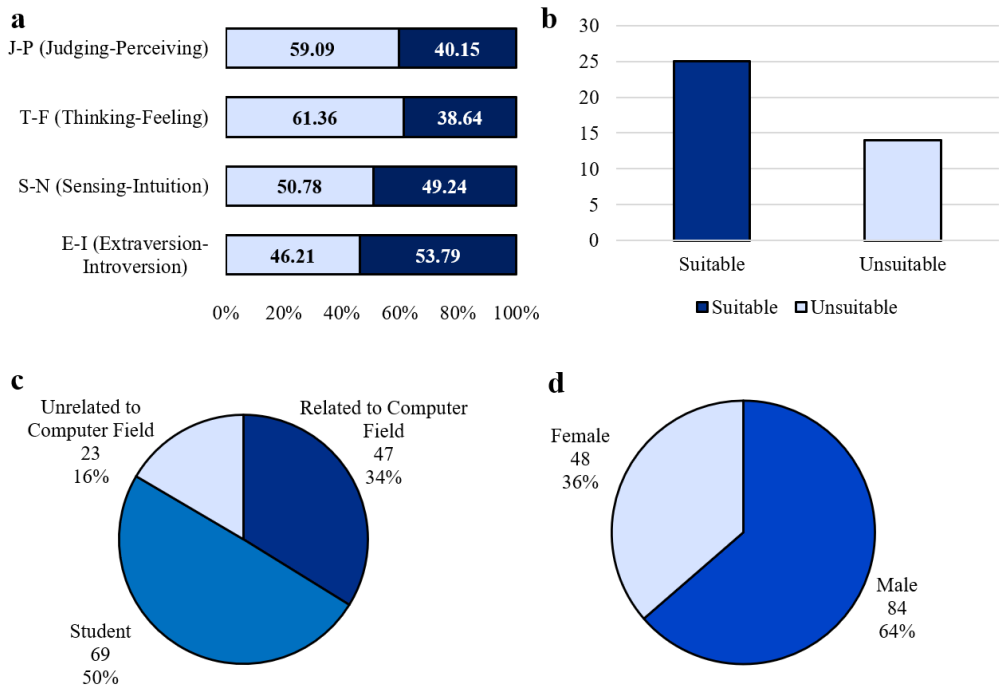

图 4. 第二次实验参与者的 MBTI 人格维度和人口统计特征。(a) 四个 MBTI 二分维度的百分比分布。(b) 根据证据和专家判断，被归类为适合或不适合计算机行业岗位的参与者人数。(c) 参与者背景分布。(d) 参与者的性别分布。

本阶段评估了多种监督学习算法。由于逻辑回归具有可解释性，且在处理小型数据集时表现稳健，因此被选为基线模型。此外，还采用了随机森林分类器等基于集成的方法，以捕捉人格维度与适配结果之间可能存在的非线性关系。研究还探索了支持向量机，因为它们适用于样本有限的二元分类问题，并能有效处理高维特征空间。此外，还纳入了梯度提升方法，以评估序列集成学习能否通过迭代纠正分类错误来提升预测性能。

鉴于已标记实例数量相对较少，在整个阶段中，控制过拟合都是首要考虑因素。为解决这一问题，模型训练和评估过程中采用了交叉验证。数据集被划分为多个折，每个模型都在不同的数据划分上进行训练和测试，以确保性能估计稳定，不受特定数据划分的影响。本阶段选取准确率作为主要评估指标，因为在职业推荐场景中，假阳性和假阴性预测都会产生实际影响。在各折中表现稳定的模型被保留，用于为其余参与者添加标签。

第一阶段完成后，数据集包含一套完整的标签，用于表示推断得出的参与者是否适合软件开发。这一经过扩充的数据集为第二个也是更关键的建模阶段奠定了基础，该阶段旨在仅使用源自游戏过程的行为特征来预测适合度。本阶段构成了本研究的核心方法学贡献，因为它评估了游戏过程中捕捉到的隐式行为信号能否可靠地取代或补充传统人格评估。

在模型训练之前，我们对行为数据集进行了广泛的预处理和特征工程。尽管前期预处理步骤已经移除了低方差特征和冗余特征，但剩余行为变量的高维性仍带来了额外挑战。特征尺度存在显著差异，反映了动作计数、时间测量值和二元指标之间的不同。为减轻尺度差异对模型性能的影响，我们采用了归一化技术，将特征重新缩放

到可比较的范围内。对于基于距离的算法和基于梯度的优化方法而言，这一步尤为重要。

特征选择和降维在这一阶段的方法论中发挥了核心作用。鉴于行为变量数量远多于已标记实例数量，降低维度对于提升模型泛化能力和可解释性至关重要。研究探索了多种特征选择策略。采用单变量统计方法评估每个特征与目标标签之间的独立关系，从而识别出具有较强判别能力的变量。将递归特征消除与监督学习模型结合，根据模型性能迭代移除信息量较低的特征。同时，利用从集成模型中提取的基于树的特征重要性度量，识别出对预测准确率贡献最大的变量。

除特征选择外，还研究了降维技术，以紧凑的形式捕捉潜在行为模式。采用线性方法将原始特征空间映射为更低维的表示，同时保留类别可分性；并探索非线性方法，以捕捉特征之间的复杂关系。随后，将这些降维后的表示作为分类模型的输入，从而评估抽象的行为嵌入相较于原始特征或经过选择的特征是否能够提升预测性能。

此阶段的模型训练采用分层K折交叉验证，以确保各折中的类别分布得以保留。鉴于存在类别不平衡，这种方法尤为重要，因为被标记为不适合的参与者人数少于被标记为适合的人数。为进一步应对这种不平衡，我们分别在对少数类进行随机过采样和不进行随机过采样的情况下开展了实验。过采样仅应用于训练折，以防止信息泄漏并维护评估结果的完整性。

我们针对不同的特征子集、降维策略和采样配置，训练并评估了多个分类模型。我们使用准确率、精确率、召回率和F1分数评估模型表现，从而全面衡量模型在不同错误类型下的效能。除定量指标外，我们还进行了定性分析，以评估模型稳定性、对特征选择的敏感度，以及在各验证折中的一致性。我们尤其关注过拟合迹象，例如训练表现与验证表现之间存在较大差异，或模型过度依赖少量特征子集。

这一建模阶段的成果是确定了一个或多个模型，这些模型仅依据游戏行为预测软件开发适配性时表现稳健且稳定。这些模型构成了所提出的基于游戏的职业推荐框架的基础，并为以下假设提供了实证支持：隐式行为评估可作为显式人格测试的一种可行替代方案。

3.6 模型评估、选择与稳健性分析

模型开发与训练完成后，最后阶段着重于评估、选择与稳健性分析。此阶段不仅旨在找出表现优异的模型，也旨在确保其稳定性、泛化能力及方法论的严谨性。鉴于该系统旨在用于职业指导，评估时同时考虑了量化表现与模型的定性行为。模型评估采用了多种互补指标。准确率用于衡量整体分类正确率，但由于可能存在类别不平衡，还辅以精确率、召回率和F1分数。精确率用于评估正向适配性预测的信度，召回率用于衡量识别真正适合个体的能力，而F1分数则综合这两方面，形成单一且易于解读的衡量指标。

4. 结果

4.1 使用基于人格的特征补全适配性标签

第一个建模阶段旨在利用迈尔斯-布里格斯问卷得出的人格维度以及补充的自我评估特征，补全缺失的适合度标签。研究选取了39名参与者

其软件开发适合度已得到确认，并将其作为真实标注。其余记录则被视为未标注样本。该任务被设定为二元分类问题，并采用k折交叉验证进行评估。所评估的机器学习算法的性能比较结果见表2。

表2. 基于个性的适合度标注的模型性能

算法	F1分数	准确率	精确率	召回率
随机森林	0.76	0.80	0.76	0.76
MLP-64h	0.76	0.80	0.76	0.76
MLP-22-63h	0.76	0.80	0.76	0.76
SVM	0.67	0.75	0.67	0.67
逻辑回归	0.41	0.70	0.35	0.50
GBM	0.33	0.50	0.33	0.33

如表2所示，随机森林和多层感知机取得了最高准确率，均为0.80，且精确率和召回率均衡（均为0.76）。这些模型在各验证折中的表现稳定。支持向量机和逻辑回归的准确率较低，而梯度提升机的表现最弱。

鉴于本阶段旨在补全数据集标签，而非优化最终的预测模型，因此准确率被视为最重要的指标。基于这一标准，最终选择随机森林来推断其余参与者的适合度标签。

为评估减少输入变量的数量是否能够提升性能，对六个基于人格的特征与适合度标签进行了相关性分析。所得相关矩阵如图5所示。将特征集缩减为相关性最高的四个变量并未提升准确率或模型稳定性。对训练指标和验证指标的检查证实，所选模型未出现过拟合。

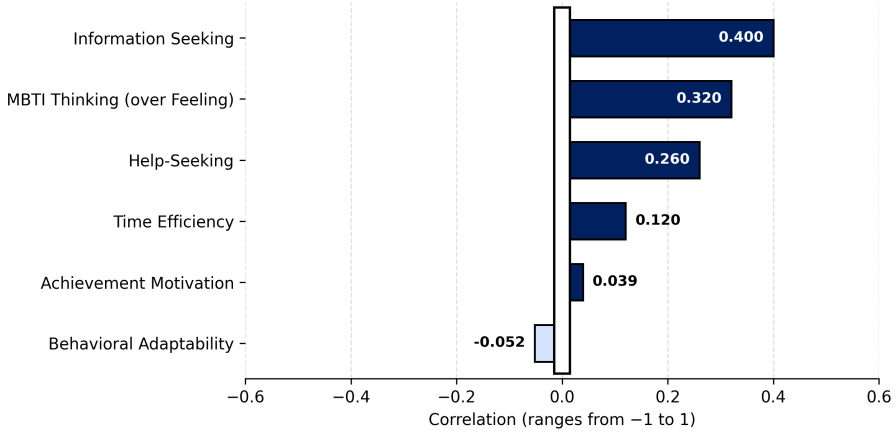

图5. 人格特征与适合度标签之间的相关性

4.2 使用游戏过程数据预测软件开发适合度 a

在第二个建模阶段，仅根据源自游戏过程的行为特征预测软件开发适配性，使用第一阶段生成的标签，并排除基于问卷的数据或人格数据。该阶段评估了游戏过程中捕捉到的隐式行为信号是否足以支持可靠预测。为探讨这一问题，研究考察了两种互补的建模策略：特征选择

和特征降维。表3所总结的特征选择结果显示，随机森林模型始终表现出色，精确率最高达0.94，准确率最高达0.88。过采样在某些情况下提高了召回率，但并未持续提升整体准确率。过采样的影响及各算法的性能如图6所示。

表3。使用特征选择进行基于游戏玩法的适用性预测

算法	特征选择	#特征 精确率	准确率		召回率	F1分数
随机森林	基于相关性	–	0.94	0.88	0.67	0.72
随机森林	单变量特征选择	10	0.94	0.88	0.67	0.72
随机森林	随机森林	5	0.94	0.88	0.67	0.72
随机森林	随机森林	15	0.94	0.88	0.67	0.72
MLP	递归特征消除 15		0.94	0.88	0.67	0.72
逻辑回归	单变量特征选择	10	0.92	0.85	0.58	0.60
随机森林过采样 递归特征消除 5			0.92	0.85	0.58	0.60
梯度提升机过采样	递归特征消除 10		0.80	0.89	0.80	0.80
逻辑回归	随机森林	5	0.77	0.85	0.65	0.68
随机森林	随机森林	10	0.77	0.85	0.65	0.68
随机森林过采样 随机森林		15	0.77	0.85	0.65	0.68
多层感知机 过采样	递归特征消除 15		0.77	0.85	0.65	0.68

特征降维结果见表4，其中测试了 LDA、PCA、t-SNE 和 UMAP 与多种分类器的组合，并比较了过采样与未过采样的情况。总体表现最佳的是 LDA 与 MLP 分类器结合过采样，达到 97%精确率和 94%准确率。非线性降维方法（t-SNE 和 UMAP）的准确率较低，结果也不够稳定。过采样的影响及算法比较见图6。

表4. 使用特征降维进行基于游戏玩法的适用性预测

算法	特征降维	精确率	准确率	召回率	F1分数
多层感知机 过采样	LDA	0.97	0.94	0.84	0.88
逻辑回归	LDA	0.95	0.91	0.75	0.81
随机森林	LDA	0.95	0.91	0.75	0.81
GBM	LDA	0.95	0.91	0.75	0.81
MLP	LDA	0.95	0.91	0.75	0.81
随机森林过采样	LDA	0.95	0.91	0.75	0.81
梯度提升机过采样	LDA	0.95	0.91	0.75	0.81
梯度提升机过采样	PCA	0.82	0.88	0.73	0.76
逻辑回归过采样	LDA	0.80	0.88	0.80	0.80
MLP	t-SNE	0.58	0.79	0.55	0.55
多层感知机过采样	T-SNE	0.54	0.78	0.53	0.53
多层感知机 过采样	UMAP	0.53	0.67	0.54	0.53

总体而言，两个建模阶段表明，基于人格的特征可以补全缺失的适合度

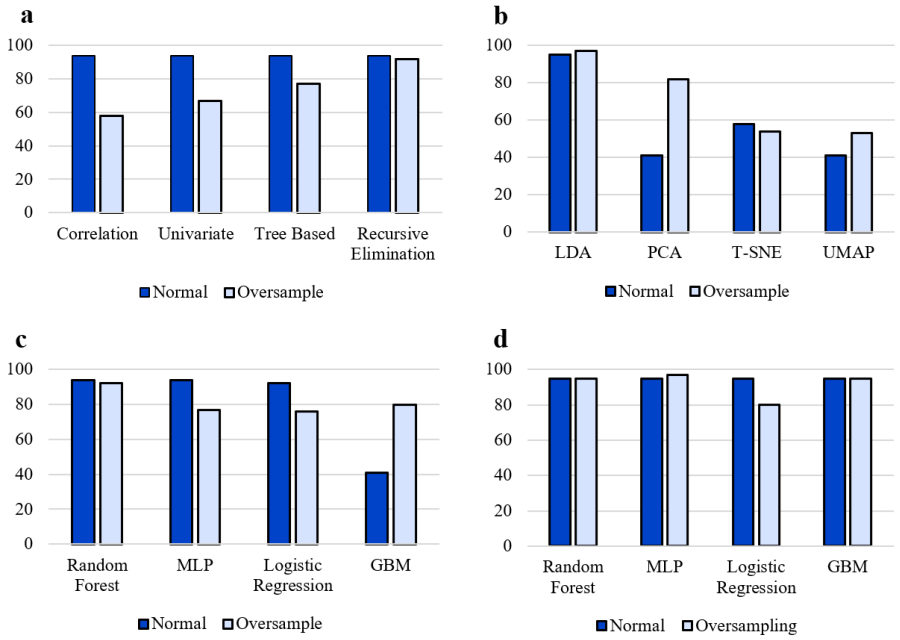

图6. 特征选择与特征降维结果的比较：(a) 特征选择方法，展示各方法在正常训练和过采样条件下表现最佳的分类器。(b) 特征降维技术，展示与各降维方法相关联的性能最高的分类器，并比较正常训练与过采样条件下的结果。(c) 特征选择方法的算法维度比较，展示各分类器在所选特征子集上的汇总表现，分别对应正常和过采样设置。(d) 特征降维方法的算法维度比较，展示分类器在降维后的特征表示上训练时的表现，并比较过采样与未过采样的情况。

标签的准确率达到 80%（表2）；仅凭从游戏玩法中提取的行为特征，就能以最高 97%精确率和 94%准确率预测软件开发适合度（表4），为严肃游戏行为数据在适配度预测中的有效性提供了有力的量化证据。

5. 讨论

在构建最终的基于游戏过程的预测模型之前，我们进行了探索性相关分析，以考察源自游戏过程的行为特征与此前识别出的四种开发者相关特质之间的关系。为确保可解释性，仅保留绝对值大于 0.25 的相关性。如图7所示，搜索相关特质越强的人，在游戏开始前越倾向于仔细查看教程。

思考导向特质与较少的记录事件、记忆型关卡中较少的猜测次数以及解谜类游戏中较高的成功率相关，但与快节奏射击类关卡中较弱的表现相关，这一点体现在较高的事件数量上。更愿意寻求帮助的参与者在游戏过程开始前更深入地学习了教学内容。较强的时间管理能力与较少的总体游戏事件以及在射击类游戏中表现提升相关。

最后，推断出的软件开发者适合度标签与游戏行为之间的相关性（图7）表明，适合度较高的参与者在益智类游戏中获胜次数更多、完成的支线挑战更多、浏览菜单的频率更高，且暂停、重试和放弃操作更少。

多层感知器（MLP）模型表现出色，尤其是在结合线性判别分析（LDA）和过采样时；这并非偶然，而是反映了模型架构与数据结构之间的良好匹配。

首先，游戏数据集具有高维、连续的特点，包含大量相互依赖的行为指标，例如动作频率、时间度量和交互模式。这些特征并不

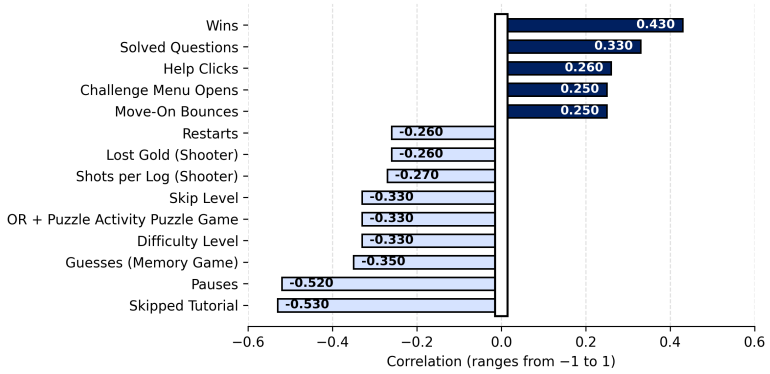

图7. 游戏过程行为与适合度标签之间的相关性

遵循与适合度标签之间简单的线性关系。多层感知机非常适合这种情况，因为它们能够对特征的非线性组合建模，并捕捉难以用线性分类器表达的交互作用。

其次，将线性判别分析（LDA）用作特征降维步骤，对稳定多层感知机（MLP）起着关键作用。LDA 将原始行为特征空间投影到一个低维表示中，从而明确地最大化类别可分性。这种转换在保留最具判别性的行为信息的同时，减少了噪声和冗余。因此，MLP 能够在紧凑、结构良好的输入空间中运行，从而缓解过拟合并提升泛化能力；考虑到数据集的规模，这一点尤为重要。

第三，与其他模型的比较进一步支持了使用 MLP 的合理性。随机森林模型表现良好，并取得了有竞争力的准确率和精确率，这表明集成方法也能够捕捉相关的行为模式。然而，随机森林依赖轴对齐划分，在连续空间中可能难以表示平滑的决策边界。相比之下，多层感知机能够学习平滑的分布式表示；当行为特质是由多个微弱信号组合而成，而非由少数主导特征决定时，这一优势尤为明显。

最后，t-SNE 和 UMAP 等非线性降维技术表现较差，这进一步印证了所选方法的适用性。虽然这些方法在可视化方面效果出色，但在监督学习情境下，它们无法保留类别判别结构。基于 LDA 的表示始终优于这些替代方法，这表明，游戏数据中的判别信息最适合通过监督线性投影后接非线性分类来捕捉，而这正是 LDA ＋ MLP 流程所采用的配置。

对行为结果和建模结果的讨论凸显了本研究的一项关键意义：基于游戏的评估能够推动评价方式从自我报告特质转向观察到的能力。严肃游戏通过记录个体在受到限制时如何规划、适应、坚持并确定优先事项，使研究者能够了解难以伪装或有意识操纵的决策层面的行为。

从实践角度来看，这意味着基于游戏的系统可作为一种可扩展、富有吸引力且偏差更小的工具，用于早期职业指导或候选人筛选。重要的是，结果表明，此类系统无需依赖明确的人格问卷——这类问卷容易受到自我报告偏差和疲劳的影响——而是可以通过游戏过程隐式推断相关特质。

6. 局限性与未来工作

尽管本研究取得了令人鼓舞的成果，但在解读研究发现时仍须承认若干局限性，而这些局限性也为未来研究指明了重要方向。

一个主要局限在于，具有经过可靠验证的软件开发适合度标签的参与者子集相对较小。尽管完整数据集由132名有效参与者组成，但只有39人依据已确认的学术或职业背景被标记为适合或不适合。这些标签

构成了监督学习过程的基础。尽管采用了交叉验证和多种建模策略来减少过拟合，但已标记子集的规模有限，仍可能给所学习的决策边界带来不确定性。未来的工作应致力于扩大经过验证的标签样本，尤其是通过与行业合作伙伴开展合作或对职业发展结果进行纵向跟踪。

适合度标签是根据自我报告的背景信息和专家判断综合得出的，而非直接测量长期工作表现或职业成功。尽管这种方法反映了职业指导早期研究面临的现实限制，但也会引入潜在的标签噪声。部分被归类为适合的参与者最终可能无法胜任专业软件开发岗位，反之亦然。未来研究可通过纳入客观绩效指标来增强效度，例如随着时间推移收集学业成绩、编程评估、实习评价或职场绩效评估。此外，未来工作还可探索采用回归模型而非分类方法，来估算个体对特定工作的适合程度。此类模型可提供连续的适合度评分，更细致地洞察个体的适合程度，并识别可通过提升特定技能来增强其就业准备度的领域。

此外，将游戏过程扩展到一天中的不同时段以及连续多天，对参与者进行评估，可进一步提高结果的稳健性。人类表现和行为会受到情绪、疲劳和认知状态等波动因素的影响；因此，在不同时间条件下通过游戏或问卷收集数据，能够更可靠、更具代表性地评估个体特征。

尽管每个游戏阶段都旨在针对特定的认知或行为构念，但无法保证每种游戏机制都只测量其预期特质。例如，谜题类关卡中的表现可能不仅反映规划能力，也会受到玩家此前接触类似游戏的经历或抽象推理技能的影响。同样，限时关卡可能同时涉及压力耐受力、运动技能以及对动作游戏的熟悉程度。多阶段设计通过对不同任务中的行为进行三角验证，有助于缓解这一问题；未来研究还可纳入正式的构念效度研究或受控消融实验，以更好地分离各项特质的独立贡献。

线性判别分析与多层感知器模型相结合，取得了最出色的预测效果。尽管神经网络模型效果良好，但对于哪些具体行为模式促成了各项个体预测，其透明度有限。在职业指导或人员筛选等高风险情境中部署此类模型时，可解释性至关重要。未来研究应结合可解释机器学习技术（如 SHAP 或特征归因分析），以提升透明度和用户信任。

所提出的框架专为评估个体是否适合从事软件开发而设计。尽管所捕捉的许多行为特质（如规划、适应能力和问题解决能力）也与其他技术职业相关，但游戏机制和建模流程均针对这一领域进行了优化。若要将该框架扩展至其他职业路径，则需要重新设计某些阶段，并重新校准特质映射。这为未来的拓展提供了一个有前景的方向。最后，要确定这些工具的信度及其在现实环境中的适用性，还需要在受控或临床环境下进行长期评估。本研究并不声称提出了最终或一劳永逸的解决方案；相反，它迈出了一步，旨在评估间接方法和基于行为的方法在心理评估中的可行性。此外，这些工具旨在帮助个人找出与其当前特征相符的职业选择，而非强制或规定具体的职业决定。

7. 结论

本研究探讨了使用严肃游戏行为数据预测个体是否适合软件开发岗位的可行性，以此作为传统基于问卷的人格评估的替代方案。研究结合迈尔斯-布里格斯框架的洞察与经实证验证、针对开发者的人格特质，设计了一款多阶段游戏，在受控而不失趣味的条件下引发有意义的认知与行为信号。实验结果有力地证明，源自游戏过程的行为特征可作为预测软件开发适配性的可靠指标，即使缺乏明确的人格数据，也能达到较高的准确率和精确率。

从实践角度来看，该框架可支持早期职业指导、教育咨询和探索性筛查，通过提供富有吸引力、低压力的评估方式，尊重用户自主权，且不预设职业发展结果。本研究并未声称能够提供最终的、经临床验证的甄选工具；相反，它为基于游戏的行为评估奠定了方法论基础，并提供了实证依据。

披露声明

作者声明，就本文的发表而言不存在利益冲突。

资金支持

本研究未获得来自公共、商业或非营利部门资助机构的任何专项资助。

参考文献

[1] Landers, R. N. 和 Sanchez, D. R.，《用于员工选拔的基于游戏、游戏化和游戏化设计的评估：定义、区别、设计与验证》，International Journal of Selection and Assessment 30(1)，1–13（2022）。[2] ST˘ANESCU, D. F.、IONIT˘A, C. 和 IONIT˘A, A.-M.，《人员招聘与选拔中的游戏思维：优势与劣势》，《后现代开放》11(2)（2020）。[3] Amin, N. A. M.、Sjarif, N. N. A. 和 Yuhaniz, S. S.，《使用 MTCNN、RetinaFace 和 Dlib 人脸检测器进行面部特征提取以识别人格特质的比较研究》，《马来西亚计算机科学杂志》38(2)，22–38（2025）。[4] Acu˜na, S. T.、G´omez, M. 和 Juristo, N.，《人格、团队过程和任务特征如何与工作满意度及软件质量相关？》，《信息与软件技术》51(3)，627–639（2009）。[5] VarastehNezhad, A.、Agahi, B.、Elyasi, S.、Tavasoli, R. 和 Farbeh, H.，《荣格认知功能如何解释计算机行业职业中的 MBTI 类型分布》，arXiv 预印本 arXiv:2504.17248（2025）。[6] Rodr´ıguez Montequ´ın, V.、Mesa Fern´andez, J. M.、Balsera, J. V. 和 Garc´ıa Nieto, A.，《使用 MBTI 评估项目式学习中工程团队的成功》，《国际技术与设计教育杂志》23(4)，1127–1146（2013）。[7] Vinciarelli, A. 和 Mohammadi, G.，《人格计算综述》，《IEEE 情感计算汇刊》5(3)，273–291（2014）。[8] Kedar, S. 和 Bormane, D.，《自动人格评估：系统综述》，载于 [*2015* 国际信息处理会议（ICIP）], 326–331，IEEE（2015）。[9] Pittenger, D. J.，《关于迈尔斯-布里格斯类型指标的警示性评论》，《咨询心理学杂志：实践与研究》57(3)，210（2005）。[10] Elyasi, S.、Taghiyareh, F. 和 Varaste, N. A.，《专业软件工程师的 MBTI 人格类型与特质》，（2023）。[11] Randall, K.、Isaacson, M. 和 Ciro, C.，《迈尔斯-布里格斯人格类型指标的效度与信度：系统综述与荟萃分析》，Journal of Best Practices in Health Professions Diversity 10(1)，1–27（2017）。[12] McCord, J.-L.、Harman, J. L. 和 Purl, J.，《类游戏人格测试：一种新兴的人格评估模式》，Personality and Individual Differences 143，95–102（2019）。[13] Schweiger, D. M.，《测量管理者认知风格：论迈尔斯-布里格斯类型指标的逻辑效度》，《商业研究杂志》13(4)，315–328（1985）。[14] Harman, J. L.，《游戏化人格评估可减少作假和粗心作答》，载于 [人因与工效学学会年会论文集], 10711813251363208, SAGE 出版社，SAGE CA：洛杉矶，CA（2025）。[15] Motlagh, S. M. H.、Rezvani, M. H. 和 Khounsiavash, M.，《用于识别人格特质的人工智能方法：系统综述》，《神经计算》，130301（2025）。[16] VarastehNezhad, A.、Tavasoli, R.、Elyasi, S.、LotfiNia, M. 和 Farbeh, H.，《心理健康领域的人工智能：大语言模型针对抑郁、焦虑和压力问题回答的情绪与情感分析》，arXiv 预印本 arXiv:2508.11285（2025）。

[17] Naz, A.、Khan, H. U.、Bukhari, A.、Alshemaimri, B.、Daud, A. 和 Ramzan, M.，《用于人格特质检测的机器学习与深度学习：全面综述及开放性研究挑战》，Artificial Intelligence Review 58(8)，239（2025）。[18] Elyasi, S. 和 Taghiyareh, F.，《通过引入益智游戏开展基于 MBTI 的人格评估》，载于[*2023 9*届网络研究国际会议（ICWR）], 102–107，IEEE（2023）。[19] Leutner, F.、Codreanu, S.-C.、Brink, S. 和 Bitsakis, T.，《招聘中基于游戏的认知能力评估：效度、公平性与应试体验》，Frontiers in Psychology 13，942662（2023）。[20] Elyasi, S.、VarastehNezhad, A. 和 Taghiyareh, F.，《从游戏到预测：利用机器学习模型评估玩家行为中的抑郁与焦虑》，International Journal of Serious Games 12(1)，83–102（2025）。[21] van Nimwegen, C.、van Oostendorp, H.、Modderman, J. 和 Bas, M.，《GameDNA 的测试案例：将用于测量人格特质的严肃游戏概念化》，载于 [*2011 16*届计算机游戏国际会议（CGAMES）], 217–222，IEEE（2011）。[22] Zulkifly, A. N.，《通过电子游戏开展人格评估》，博士论文，塔斯马尼亚大学（2019）。[23] Haizel, P.、Vernanda, G.、Wawolangi, K. A. 和 Hanafiah, N.，《基于五因素模型的人格评估电子游戏》，Procedia Computer Science 179，566–573（2021）。[24] Graham, L. T. 和 Gosling, S. D.，《与不同魔兽世界游戏动机相关的人格特征画像》，Cyberpsychology, Behavior, and Social Networking 16(3)，189–193（2013）。[25] Wohn, D. Y. 和 Wash, R.，《带有线索的虚拟“房间”：通过城市模拟游戏中的空间定制检测人格》，Computers in Human Behavior 29(1)，155–159（2013）。[26] Feldman, J.、Monteserin, A. 和 Amandi, A.，《探索利用在线电子游戏检测人格二分类型》，Online Information Review 41(5)，598–610（2017）。[27] M¨uller, J. 和 Schwieren, C.，《信任博弈中的大五人格因素》，Journal of Business Economics 90(1)，37–55（2020）。[28] Vollmer, C.、Randler, C.、Horzum, M. B. 和 Ayas, T.，《青少年的电脑游戏成瘾及其与时间型和人格的关系》，Sage Open 4(1)，2158244013518054（2014）。[29] Bean, A. 和 Groth-Marnat, G.，《电子游戏玩家与人格：理解游戏风格的五因素模型》，Psychology of Popular Media Culture 5(1)，27（2016）。[30] Seok, S. 和 DaCosta, B.，《预测电子游戏行为：考察人格与手机游戏体验之间的关系》，Games and Culture 10(5)，481–501（2015）。[31] Worth, N. C. 和 Book, A. S.，《电子游戏行为的维度及其与人格的关系》，Computers in Human Behavior 50，132–140（2015）。[32] Cruz, S.、Da Silva, F. Q. 和 Capretz, L. F.，《软件工程中人格研究的四十年：一项映射研究》，Computers in Human Behavior 46，94–113（2015）。[33] Furnham, A.，《大五人格与大四人格之比较：迈尔斯-布里格斯类型指标（MBTI）与 NEO 人格量表五因素人格模型之间的关系》，Personality and individual differences 21(2)，303–307（1996）。[34] Capretz, L. F.，《软件工程师真的是工程师吗？》，World Transactions on Engineering and Technology Education 1(2)，233–235（2002）。[35] Barends, A. J. 和 de Vries, R. E.，《模拟选拔情境中人格评估游戏的构念效度，以及识别标准的能力和性格洞察力的调节作用》，International Journal of Selection and Assessment 31(1)，120–134（2023）。[36] Wear, H.、Sanchez, D. R. 和 Garcia, K.，《概括基于游戏的评估（GBAs）：使用人格评估游戏推断职业兴趣的价值》，[37] Gao, M.、Wang, J. 和 Yang, J.，《基于关联规则挖掘研究电子游戏中人格与行为之间的关系》，Mathematics 11(3)，772（2023）。[38] Quwaider, M.、Alabed, A. 和 Duwairi, R.，《利用五因素模型特质和机器学习预测人格的射击类电子游戏》，Simulation Modelling Practice and Theory 122，102665（2023）。[39] Van Lankveld, G.、Spronck, P.、Van den Herik, J. 和 Arntz, A.，《作为人格画像工具的游戏》，载于 [*2011* IEEE 计算智能与游戏会议（CIG’11）], 197–202，IEEE（2011）。

[40] Afroza, A., Murray, K., W¨unsche, B. C., 和 Denny, P., 《Who am i?-development and analysis of an interactive 3d game for psychometric testing》, 载于 [2021年澳大拉西亚计算机科学周多会议论文集], 1–10（2021）。[41] Harman, J. L. 和 Brown, K. D., 《Illustrating a narrative: A test of game elements in game-like personality assessment》, 《International Journal of Selection and Assessment》30(1), 157–166（2022）。[42] Harman, J. L. 和 Purl, J., 《Advances in game-like personality assessment》, 《心理学趋势》32(4), 1445–1459（2024）。[43] Yudina, I., Kozlovskii, P., Pavlikova, N., Kochkina, K. 和 Sataev, P., 《Analysis of possibilities of using game statistics of the cloud quest in assessment of personality》, 载于 [国际交互式协作学习会议], 813–822, 施普林格（2021）。[44] Liapis, G., Zacharia, K., Rrasa, K., Liapi, A. 和 Vlahavas, I., 《Modelling core personality traits behaviours in a gamified escape room environment》, 载于 [欧洲游戏化学习会议], **16(1)**, 723–731（2022）。[45] Wu, F. Y., Mulfinger, E., Alexander III, L., Sinclair, A. L., McCloy, R. A. 和 Oswald, F. L., 《Individual differences at play: An investigation into measuring big five personality facets with game-based assessments》, 《International Journal of Selection and Assessment》30(1), 62–81（2022）。[46] Miller, L. C., Jeong, D. C. 和 Christensen, J. L., 《Virtual environments for the representative assessment of personality: Ve-rap》, 载于 [测量与建模个体及情境], 223–252, 爱思唯尔（2021）。[47] Mayers, R. B., 《人格特质与电子游戏角色选择之间的关系》, 硕士论文, 路易斯安那州东南大学（2021）。[48] Yu, S., 《警告对游戏化人格评估的影响》（2021）。[49] Palhano, D. B., Machado, L. d. S. 和 De Almeida, A. A. F., 《Paki mirabolandia：一款用于识别玩家人格的严肃游戏》, 载于 [*2020* IEEE第八届严肃游戏与健康应用国际会议（SeGAH）], 1–5, IEEE（2020）。[50] Waters Jr, L. J., Garcia, K., Rueda, A., Waters, L., Sanchez, D., Garcia, K. 和 Rueda, A., 《探索基于游戏的人格评估中的伪造分数》, 未发表手稿*]*。旧金山州立大学（2022）。[51] Wirth, M., Gradl, S., Mehringer, W. A., Kulpa, R., Rupprecht, H., Poimann, D., Laudanski, A. F. 和 Eskofier, B. M., 《在虚拟现实中评估团队运动员的人格特质》, 载于 [*2020* IEEE虚拟现实与3D用户界面会议摘要与研讨会（VRW）], 101–108, IEEE（2020）。[52] Ponticorvo, M., Ferrara, F., Mugione, M., Barbaranelli, C. 等, 《大五人格游戏：一款用于评估和训练学生掌握大五理论的数字角色扮演游戏》, 载于 [PSYCHOBIT], （2020）。[53] Raja, H. H., Patel, V. K., Tiwari, D. S., Kadavala, N. C. B. 和 Patel, N., 《游戏成瘾：卫生专业本科生游戏特征与人格特质研究》, 《印度精神病学年鉴》4(2), 164–169（2020）。[54] Wang, Z., Sapienza, A., Culotta, A. 和 Ferrara, E., 《基于角色的在线游戏中的人格与行为》, 载于 [*2019* IEEE游戏会议（CoG）], 1–8, IEEE（2019）。[55] Olson, A. 和 Siraj, P., 《根据电子游戏中展现的人格预测工作绩效》（2019）。[56] Barnes, P., 《重新审视程序员偏执》, 载于 [第十三届SIGCPR年度会议论文集],114–131（1975）。[57] Rehman, M., Mahmood, A. K., Salleh, R. 和 Amin, A., 《将软件工程师的工作要求映射到大五人格特质》, 载于 [*2012* 计算机与信息科学国际会议（ICCI）],2, 1115–1122, IEEE（2012）。[58] Rasch, R. H. 和 Tosi, H. L., 《影响软件开发人员绩效的因素：一种综合方法》, 《MIS季刊》, 395–413（1992）。[59] Martin, C. R., 《 [认识类型：基础知识], 》, 心理类型应用中心（1997）。[60] Schweda Nicholson, N., 《口译员受训者的人格特征：迈尔斯-布里格斯类型指标（MBTI）》（2005）。[61] Amin, A., Basri, S., Rehman, M., Capretz, L. F., Akbar, R., Gilal, A. R. 和 Shabbir, M. F., 《人格特质与知识收集行为对程序员创造力的影响》, 《信息与软件技术》128, 106405（2020）。

[62] Vadlamani, S. L. 和 Baysal, O.，《研究软件开发人员在 Stack Overflow 和 GitHub 上的专业能力与贡献》，载于 [*2020 IEEE*软件维护与演进国际会议（*ICSME*）],312–323，IEEE（2020）。[63]RAJAKARTHIKEYAN, K., BALASUBRAMANIAN, P., MANIVANNAN, L. 和 SHANMUGASUN- DARAM, M. R.，《班加罗尔部分IT公司软件开发人员的特征与能力研究》[64] Li, P. L., Ko, A. J. 和 Begel, A.，《优秀软件工程师有何不同？》，《实证软件工程》**25(1)**，322–352（2020）。[65] Celikten, A. 和 Cetin, A.，《根据人格类型和技能为软件工程师分配产品开发角色》，《学术科学期刊》（2月）（2017）。[66] Li, P. L., Ko, A. J. 和 Zhu, J.，《怎样才能成为优秀的软件工程师？》，载于 [*2015 IEEE/ACM*第*37*届*IEEE*软件工程国际会议], **1**, 700–710，IEEE（2015）。[67] Ahmed, F., Capretz, L. F., Bouktif, S. 和 Campbell, P.，《软技能与软件开发：来自软件行业的反思》，*arXiv*预印本 *arXiv:1507.06873*（2015）。

附录A。游戏设计、机制与数据收集详情

本附录详细介绍了为行为和人格评估而开发的严肃游戏，包括其机制、阶段以及数据采集过程。

A.1 分组交换拼图游戏

第一款游戏是一种拼图游戏，玩家必须在受限网格内，在有限步数之内交换两组棋子的位置。只有在选中棋子后通过键盘输入，才能将棋子移动到有效且未被占用的单元格。无效移动将被忽略，且不计入步数。

每个关卡都有一个在时间和移动步数方面的最优解。如果时间限制到期，玩家可以认输或继续游戏。该游戏由三个阶段组成：入门教程阶段、中等难度阶段和高难度阶段。图8展示了实验中使用的分组交换拼图游戏界面。

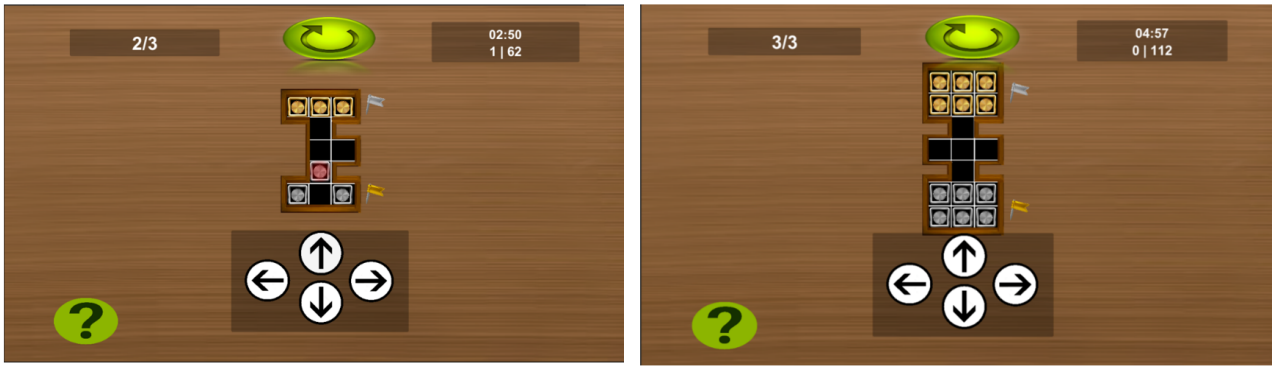

图8。分组交换拼图游戏界面

为减少挫败感，玩家可以使用数量有限的“跳过代币”，跳过拼图阶段而不受惩罚。初始提供三个跳过代币，完成可选的逻辑性强或数学挑战后，还可以获得额外的跳过代币。图9展示了跳过代币和可选挑战机制的示例。

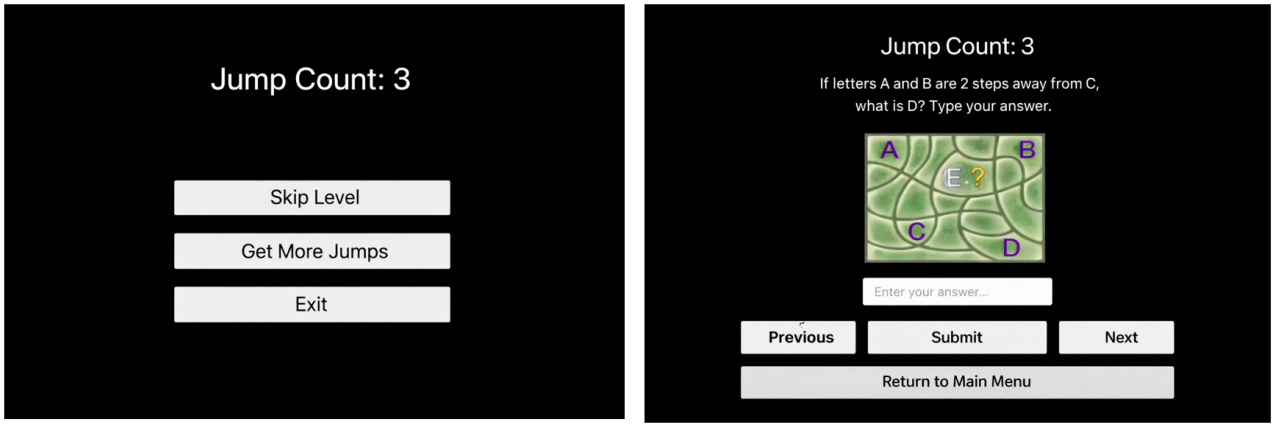

图9. 跳过令牌和可选挑战机制

A.2 记忆配对游戏

记忆游戏会向玩家展示成对的数字，持续五秒。随后，卡片会翻面，玩家必须凭记忆找出匹配的数字对。如果玩家未能成功，可以反复尝试，以鼓励他们采用高效的记忆策略。

该游戏包含三个难度等级，卡片数量逐步增加。为确保玩家充分投入基于记忆的行为，该游戏未提供跳过选项。图10展示了实验中使用的记忆配对游戏界面

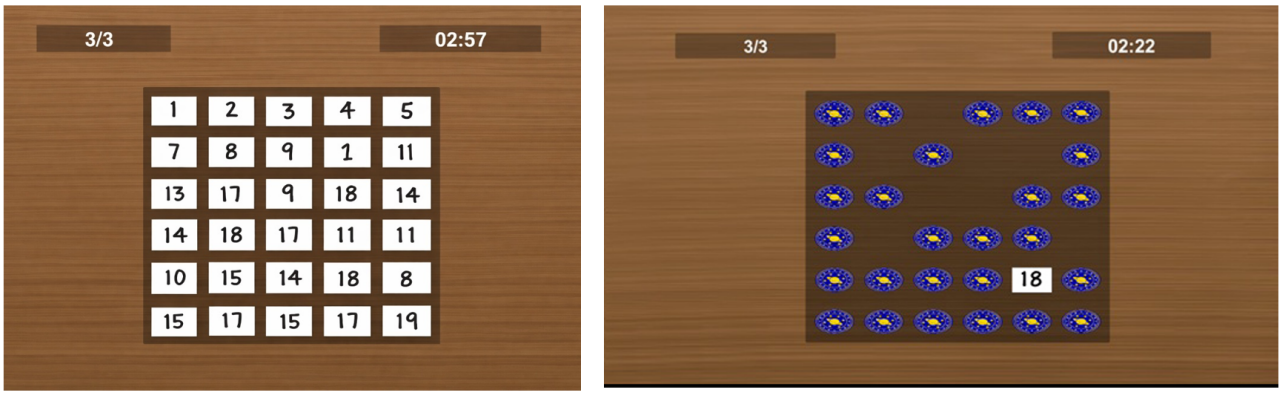

图10。记忆配对游戏界面

A.3 银河射击游戏

银河射击游戏是一款动作射击游戏，玩家必须在遭遇不同敌人类型的情况下存活120秒，其中包括攻击型、防御型和得分型实体。菜单中提供可选的支线挑战。图11展示了该射击游戏的教程界面示例。



图11。射击游戏中的教程界面

三次尝试失败后，玩家可以选择放弃并继续进行。多个教程界面会解释操作方式和敌人行为。该游戏主要捕捉与时间管理、反应速度和压力相关的决策行为。银河射击游戏的界面如图12所示。

A.4 障碍物重排路径游戏

在这个拼图中，玩家必须移动障碍物，引导目标棋子到达指定终点。障碍物的移动模式受限，包括单格方块、竖向和横向双格方块，以及正方形



图12. Galaxyshooter 游戏界面

形状方块（如图13所示）。游戏包含三个复杂度逐渐提高、移动自由度有限的阶段。

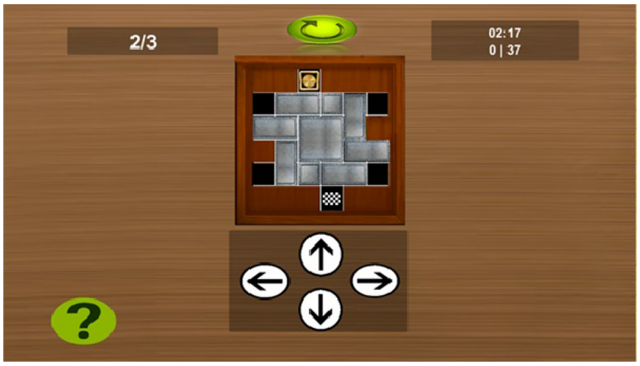

图13. 障碍重排路径游戏界面

A.5 图遍历游戏

这款游戏要求玩家遍历图中的所有顶点，且不得重复经过任何节点。玩家选择方向后，角色会自动持续移动，直到遇到先前经过的节点或障碍物。根据玩家反馈，我们简化了原本复杂的规则，最终版本包含四个难度逐步提升的阶段。图14展示了图遍历游戏的界面。

A.6 游戏过程日志记录与数据提交

完成游戏后，玩家需要将其游戏过程日志提交至服务器（图15）。系统会为每位参与者生成一个唯一的五位数追踪码，该追踪码对应于根据其游戏过程日志生成的哈希值的末五位。该代码能够在分析过程中实现匿名追踪和可靠的数据关联。

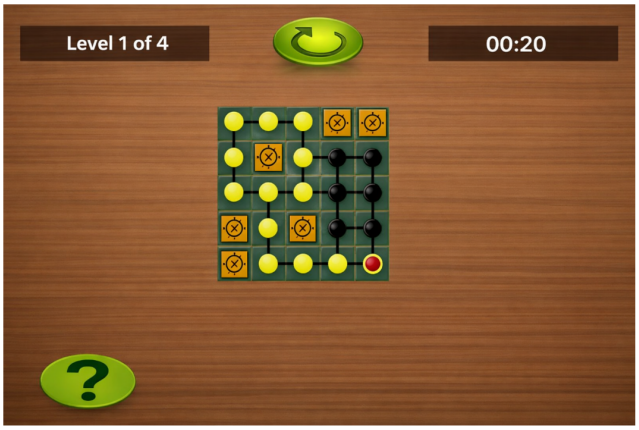

图14. 图遍历游戏界面

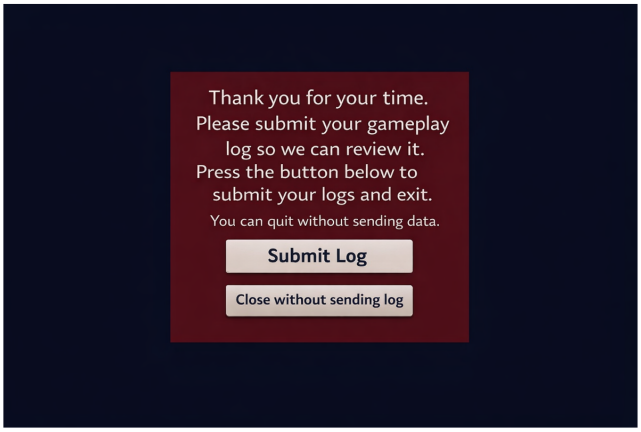

图15. 游戏数据提交与追踪界面

A.7 未纳入的原型游戏

在测试版测试期间，我们还评估了其他游戏，但最终版本未予采用。其中包括一款旨在通过叙事选择评估人格的叙事卡牌角色扮演游戏，但由于参与度较低而被移除。另一款聚焦于战略规划和资源管理的塔防游戏也被排除在外，原因是其复杂程度过高，且参与者没有足够的学习时间。这两款游戏展示于图16中。



图16. 未纳入的原型游戏设计

附录 B. 变量列表

唯一游戏过程追踪代码（字符串）- 参与者年龄（数值）- 参与者性别（枚举）- 性别标记：男性（布尔值）- 性别标记：女性（布尔值）- 计算机工程背景（布尔值）- MBTI人格类型（枚举）- MBTI外向—内向（枚举）- MBTI感觉—直觉（枚举）- MBTI思考—情感（枚举）- MBTI判断—知觉（枚举）- MBTI SJ功能组（布尔值）- MBTI SP功能组（布尔值）- MBTI NF功能组（布尔值）- MBTI NT功能组（布尔值）- 已确认的编程兴趣（布尔值）- 预测的编程适合度（布尔值）- 求助行为（布尔值）- 在线搜索问题解决技能（布尔值）- 时间管理能力（布尔值）- 竞争动机（布尔值）- 行为灵活性与适应性（布尔值）- 主要游戏平台（枚举）- 每周平均游戏时长（数值）- 所选游戏难度（枚举）- 游戏日志总数（数值）- 游戏总时长（分钟）（数值）- 游戏暂停总次数（数值）- 游戏重启总次数（数值）- 分组交换拼图游戏：日志数（数值）- 分组交换拼图游戏：游戏时长（分钟）（数值）- 分组交换拼图游戏：跳过令牌使用次数（数值）- 分组交换拼图游戏：认输操作次数（数值）- 障碍物重排路径游戏：日志数（数值）- 障碍物重排路径游戏：游戏时长（分钟）（数值）- 记忆配对游戏：日志数（数值）- 记忆配对游戏：游戏时长（分钟）（数值）- 记忆配对游戏：猜测总次数（数值）- 记忆配对游戏：正确猜测次数（数值）- 记忆配对游戏：错误猜测次数（数值）- 教程参与次数（数值）- 教程跳过次数（数值）- 测验导航交互次数（数值）- 银河射击游戏：游戏过程日志数（数值）- 银河射击游戏：达到的最高分（数值）- 银河射击游戏：总射击次数（数值）- 银河射击游戏：总损失生命数（数值）- 银河射击游戏：重启次数（数值）- 银河射击游戏：生成的强化道具（数值） - 银河射击游戏：收集的强化道具（数值） - 银河射击游戏：生成的金币（数值）- 银河射击游戏：收集的金币（数值）- 银河射击游戏：损失的金币（数值）- 银河射击游戏：爆炸损失的金币（数值）- 银河射击游戏：生成的敌人（数值）- 银河射击游戏：射击摧毁的敌人（数值）- 银河射击游戏：敌人与玩家碰撞次数（数值）- 银河射击游戏：生成的小行星（数值）- 银河射击游戏：射击摧毁的小行星（数值）- 银河射击游戏：小行星与玩家碰撞次数（数值）- 银河射击游戏：向左移动次数（数值）- 银河射击游戏：向右移动次数（数值）- 银河射击游戏：水平移动总次数（数值）- 银河射击游戏：离开左侧边界次数（数值）- 银河射击游戏：离开右侧边界次数（数值）- 银河射击游戏：离开边界总次数（数值）- 银河射击游戏：生命存活挑战已完成（布尔值）- 银河射击游戏：敌人消灭挑战已完成（布尔值）- 银河射击游戏：小行星摧毁挑战已完成（布尔值）- 银河射击游戏：金币收集挑战已完成（布尔值）- 银河射击游戏：不使用武器挑战已完成（布尔值）- 银河射击游戏：分数达成挑战已完成（布尔值）- 图遍历游戏：日志数（数值）- 图遍历游戏：游戏时长（分钟）（数值）- 银河射击游戏：每分钟游戏日志数（数值）- 银河射击游戏：每分钟胜场数（数值）- 银河射击游戏：每分钟射击次数（数值）- 银河射击游戏：每条游戏日志的射击次数（数值）- 银河射击游戏：每分钟损失生命数（数值）- 银河射击游戏：每条游戏日志的损失生命数（数值）- 银河射击游戏：强化道具收集效率比（数值）- 银河射击游戏：金币收集效率比（数值）- 银河射击游戏：敌人碰撞率（数值）- 银河射击游戏：小行星碰撞率（数值）- 银河射击游戏：方向移动偏向比（数值）- 银河射击游戏：边界离开偏向比（数值）