北京大学数字金融研究中心
Institute of Digital Finance Peking University

人工智能促进高质量充分就业系列报告之一

“时雨润无声”：AI 训练赋能劳动者职业发展

北京大学数字金融研究中心课题组

课题组成员[①]

黄益平、王靖一、费越、李振华、王芳、李勇国

技术支持团队

张韶容、钟亮华、樊雪娟、阎峰、耿东东、陈凯勋、
叶政君、邵黎明

2025 年 9 月

[①] 黄益平为北京大学数字金融研究中心主任，北京大学博雅特聘教授、北京大学国家发展研究院院长；王靖一为北京大学数字金融研究中心特约研究员，中央财经大学金融学院助理教授；费越为德国曼海姆大学助理教授；李振华、王芳和李勇国为蚂蚁集团研究院研究人员。在课题研究与报告撰写过程中，课题组得到了北京大学数字金融研究中心和蚂蚁集团各位领导和同事的大力支持和帮助，特此致谢。本报告不代表北京大学数字金融研究中心和蚂蚁集团的观点，所有问题由课题组负责。课题组感谢李昀达、姚依珣、刘曜嘉、谭励为四位同学的助研工作。

目录

一、内容提要 1
二、研究背景 2
三、研究方法 5
四、主要研究发现 5
（一）AI 训练的整体影响 5
1. 对收入的影响 6
2. 对客户满意度的影响 7
3. 对服务标准程度的影响 8
（二）分人群的 AI 训练影响 9
1. 不同性别的 AI 训练影响差异 10
2. 不同年龄段的 AI 训练影响差异 11
3. 不同所在区域的 AI 训练影响差异 13
（三）AI 训练作用机制的初步分析 14
1. AI 身份与性格设计的影响 15
2. AI 情绪设定的影响 17
五、总结与讨论 18
参考文献： 20

一、内容提要

近年来，随着以大语言模型为代表的生成式人工智能飞速发展，关于人工智能对于就业的影响成为国内外研究关注的焦点。部分研究通过招聘文本描述的分析判断岗位所受冲击的程度；另外一些研究则利用随机试验，分析在工作过程中使用 AI 工具对于工作表现的影响。而本研究则试图更进一步地挖掘基于人工智能的培训对于劳动者能力的影响：AI 不是作为工具出现在工作过程中，而是出现在训练环节增强劳动者的工作能力，这使得我们可以尝试探索人工智能如何促进劳动者技能升级，以智能手段，应对技术带来的就业冲击。

2024 年 3 月，蚂蚁集团数字蚂力上线了基于大语言模型构建的陪练智能体，这一智能体可以在训练环节中模拟真实客户行为与客服人员交互，提升客服能力。7-10 月间，业务团队进行了随机试验：部分新入职客服使用陪练智能体训练（下简称 AI 训练），部分则保持传统方式，基于此研究团队分析了 AI 训练对于劳动者后续工作表现的影响，并对其作用机制进行了初步分析。在我们的认知范围内，这应是首个关于 AI 在训练中如何影响劳动者能力的实证研究。

研究结果显示：AI 训练对于劳动者收入、客户反馈、服务标准性上均有显著正面效果，并呈现出人群特征、地理分布上的普惠性，同时初步的分析也展示出 AI 训练一项核心优势：更加精巧、贴近现实情况的设计，可以带来更为显著的训练增益。相较于传统训练方式，AI 训练对于新入职客服，在薪资提升、客户满意度、服务标准性等方面均有显著的积极影响。具体而言，AI 训练使得新入职的客服在开始工作的前六个月，单次服务平均薪酬上升了 14.02%、日均获得客户差评数下降了 29.46%，日均人工质检不合格数量下降了 29.70%。

AI 训练同时展示出了普惠性质。得益于智能体交互的真实性，AI 训练对不同性别、年龄段、所处城市线级及城乡区分的人群均带来了显著的正面效果，且组间差距很小。同时，女性在收入、男性在减少差评率、45 岁以上人群在提升服务标准性上，均获得了强于整体平均的更大改善，展现了 AI 训练的普惠性

质。这也是本轮人工智能浪潮显著特征之一：大语言模型、智能体降低了使用相关技术的门槛，为相对弱势群体提供了改善通道。

AI 训练的核心能力之一，在于陪练智能体可以为所扮演的客户赋予不同的身份、性格与多样化多阶段的情绪，使得客服在训练场景中积累更多经验，应对工作场景中的多样化问题。研究发现，当智能体具备非一般的身份、性格，具备负面、对抗性与不少于三阶段情绪时，AI 训练的增益被进一步放大，特别是当 AI 扮演暴躁、愤怒等对抗性情绪的客户时，效果尤佳：这部分经历对抗性情绪的客服，在上岗后五个月的差评数量减少 79.79%，质检不合格数量下降 49.40%。

为在深入实施“人工智能+”行动中坚持以人民为中心的发展思想，充分发挥我国数据资源丰富、产业体系完备、应用场景广阔等优势，我们建议：（1）应当在确保用户隐私、数据安全前提下，充分发挥既有产业优势与数据沉淀，利用 AI 提升产品、服务的效率与质量，提升社会整体福利水平。（2）应当直面 AI 发展对劳动者就业质量的冲击，探索利用 AI 提升、延展劳动者工作能力的路径，充分利用智能技术提升劳动者能力，对抗技术迭代压力，根本上缓解就业压力。（3）应当重视人在工作、特别是直接与人交流的工作中的独特价值，留意考量人的体验与满意度，在追求运行效率、经济效益的同时，兼顾体验等隐性指标。

二、研究背景

以大语言模型为代表的生成式人工智能的快速发展，正深刻改变就业格局。现有研究大体可以分为三条路径：一是基于岗位暴露度的测算，二是聚焦岗位需求与企业行为的实证研究，三是通过理论模型推演 AI 对劳动市场的整体冲击。

第一，岗位暴露度测算奠定了研究基础。Felten 等（2018）首次提出将 AI 能力与 O*NET 数据库中的岗位能力相匹配，以量化不同职业受 AI 影响的程度，并建立标准化指标体系。后续 Felten 等（2023）将该方法扩展至大语言模型，发现教育、客服等语言密集型行业暴露度明显上升。Eloundou 等（2023）则强调 GPT 类模型的通用性，指出约八成劳动力可能受到至少 10%的影响。张丹丹等

（2025）在此思路上进行了本土化拓展，基于大语言模型技术暴露度测算中国劳动市场的冲击程度。研究发现，在 2018 年 1 月～2024 年 5 月，中国劳动力市场上新增职位的大语言模型人工智能技术暴露度呈现降低的趋势；暴露度较高的职业主要是对受教育程度要求较高和薪资较高的白领职业，包括会计、编辑、销售及程序员等。

第二，实证研究揭示了 AI 与劳动市场的现实互动。Acemoglu 等（2022）利用美国招聘数据发现，AI 技能需求快速上升，但就业和工资增长尚无显著变化，暗示 AI 在扩散早期主要改变技能结构而非总量。Babina 等（2024）则基于员工简历数据度量企业 AI 投入，发现高 AI 投入企业表现出显著的营业额、雇员与市值增长，其增长动力主要来自产品创新而非过程效率提升。这些研究说明，AI 对劳动市场的作用具有间接性和滞后性：在短期内，劳动者面临技能错配风险，但在长期，企业扩张和创新可能创造更多就业。

第三，理论模型描绘了 AI 冲击的可能情景。Korinek 与 Suh （2024）提出“原子任务”框架，设定自动化通过逐步提高任务可替代的复杂度阈值来重塑就业。在不同自动化速度下，劳动工资可能出现增长、骤降或回升等多种路径，资本回报则大多呈上升趋势。这类模型强调了 AI 冲击的非线性和不确定性，提醒我们政策与培训干预在缓冲短期冲击、促进长期适应中的重要作用。

总体而言，三类研究构成了由微观到宏观的逻辑链条。岗位暴露度研究提供了识别冲击的工具与指标；实证研究揭示了劳动市场和企业层面的动态反应；理论模型则拓展到宏观结构与长期演化。近期的中国研究将国际方法与本土数据结合，揭示了大语言模型对中国劳动市场的差异化影响。这表明，AI 并非单向的“岗位替代者”，而是在不同情境下同时展现“替代—创造”的双重效应。

但以上研究存在一个较大的现实制约：缺乏对于劳动者个体层面真实的 AI 暴露度的直接度量（而是以工作描述、企业投资等间接方式估测），Brynjolfsson 等（2025）通过随机试验获得了劳动者层面的直接度量，并开创性地结合大规模现场部署，直接获取了劳动者的暴露度与生产率变化。他们基于一项涵盖 5179 名客户支持坐席的实地研究发现，接入生成式 AI 助手平均提升了劳动者 14%的问题解决效率，且这种提升主要集中在新手与低技能员工群体（生产率提升高达 35%），而经验丰富的高技能员工几乎未受影响 。

我们的研究则更进一步，基于个体层面的 AI 度量，探索 AI 训练对于劳动者工作表现的影响。相较于在工作中使用 AI 工具，AI 训练的作用机制更加间接，却更为重要：AI 能否、如何改变了人的能力，这一改变的效果、持续时间、作用机制如何。本研究便是利用数字蚂力上线基于大模型的客户智能体的宝贵研究素材，进行了因果识别与影响测算。

我们研究的主要对象是数字蚂力的云客服板块：不同于传统印象中的呼叫中心模式，在固定时间、固定地点招聘客服人员提供客服服务，云客服的工作人员可以自选工作地点与工作时间，灵活的工作模式、相对丰厚的报酬吸引了许多劳动者的青睐，但是如何将这些分散的“新手”训练成可以上岗的合格客服，是一个挑战。在之前，学习材料主要以文字、图片、视频等非互动形式，之后加入了基于关键词匹配的固定剧本模式，都无法将培训与材料背诵脱离开来。

2024 年 3 月 28 日，基于大模型打造的陪练智能体上线，该智能体可以不依赖繁琐的脚本配置（明确对话机器人每个步骤的对话内容与语义识别关键点，这是之前的训练脚本的主要运行方式），使用自然语义与被训练的客服交流，并理解客服的反馈，推动训练情节。5 月 17 日，智能体进一步升级，具备更加贴近真人的音色，可以模拟特殊身份、性格带来的独特表达方式，模拟冷淡、愤怒、宽容、尊重等数十种不同情绪。

在 2024 年 7 月-9 月，为了验证 AI 训练的有效性，数字蚂力团队在部分业务板块进行了随机试验，对于同一批入职的新客服，将其随机分为两组：一部分采用 AI 训练为主，另一部分采用传统训练为主。实验主要在涉及行业知识与针对性场景强化的训练中进行，对于一些基本的、共性的场景，比如用户隐私、数据安全的训练，则统一进行。下图左图汇报了实验中客服在岗前训练阶段，AI 训练占整体时长的分布直方图，我们将 AI 训练的阈值设为 50%，即如果一个客服的岗前训练超过一半以上的时间是由 AI 完成的，我们便认为他是 AI 训练的客服（实验组），否则则为控制组，之后我们将重点比较这两类人群之间的差异。如下图右图所示，最终我们包含了 512 个实验组客服和 1497 个控制组客服。

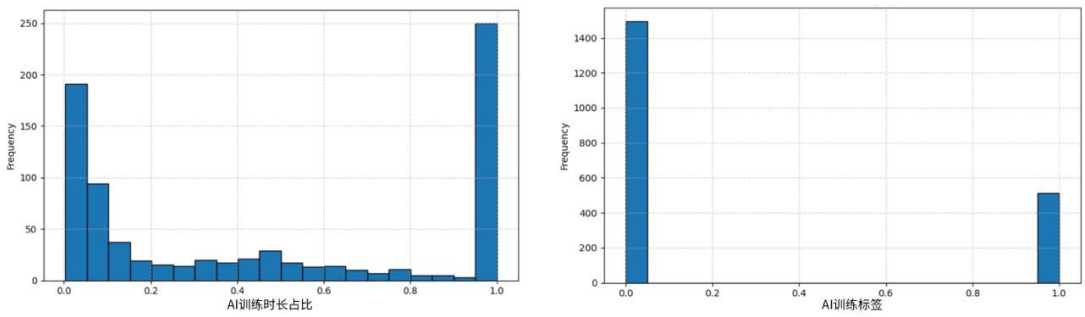

图 1 岗前 AI 训练时长占比及实验标签分布图

三、研究方法

因果森林（Causal Forest）是 Wager 等（2018）在 Athey 等（2016）的因果树（Causal Tree）的基础上的扩展，具备更强的稳健性，是一种用于评估“某项干预措施对不同人群效果是否不同”的机器学习方法，尤其适用于在存在大量个体差异的现实场景中。与传统平均处理效应（ATE）方法相比，因果森林不仅能估计整体干预效果，还能识别在哪些特定人群中效果更强（即条件平均处理效应 CATE），从而揭示政策或措施背后的异质性，这在我们的研究中有特别的价值。

该方法在本研究中尤为适用。我们评估的是 AI 训练能否提升新客服的工作表现，受试者来自不同地区、背景和技能组，表现差异显著。同时，个体培训前的属性可能同时影响培训方式的选择和最终绩效，存在“混淆因素”。因果森林在模型中纳入了多种控制变量（如性别、年龄、入职时间等），这使得我们不仅可以估计 AI 培训的总体效应，还能探索在如女性、特定技能组或经历特定培训情境的新员工中，AI 培训的优势是否更为显著，从而为研究 AI 训练的普惠性与作用机制奠定基础。

具体到本研究，我们控制了所在业务板块、所在区县、第一天工作日期、支付宝注册时间、年龄、性别、上岗前三个月支付宝支付笔数、上岗前三个月支付宝支付金额、上岗前三个月支付宝基金余额，作为分析中的协变量。

四、主要研究发现

（一）AI 训练的整体影响

在本小节，我们将汇报整体层面的分析结果，即对于客服整体而言，接受 AI 训练相较于如果没有接受 AI 训练的反事实，带来了多少增益。在不同场景下，我们分析了 AI 训练在客服上岗后首月、前两个月……直至前六个月的影响。

1. 对收入的影响

云客服的灵活性的一个集中体现便是在其工作方式与薪酬度量上，我们需要以三种不同的角度来衡量他们的薪资。在进行真正工作之前，云客服需要自主选班：每天 8-24 点的班次被以 1 个小时的颗粒度推送给客服进行选择，他们可以选取其中自己偏好的班次。故而一个客服每个月有班次的工作日、每个工作日的工作时间、工作时长都十分灵活。

而云客服的薪资组成则更为复杂，其薪酬计算方式如下：

薪水=基础薪酬*绩效系数+奖励薪酬-扣罚+其他薪酬

其中，基础薪酬=服务数量*计件薪酬+在线时长*在岗时薪。计件薪酬在同一个业务板块中的所有客服中保持统一，而在岗时薪则在更大的人群中保持相同。

绩效系数是按照客服在当月同一业务板块中所有客服的服务质量排名进行赋分，这个数字一般大于 1，高者可以超过 2.5。由于这一薪酬设计方式，片面地追求数量而忽视质量，往往不能带来更高的收益。

我们对于薪酬的考察主要关注两个维度：

小时收入：样本期内总收入/样本期内工作小时数。

单次服务收入：样本期内总收入/样本期内进行的服务数。

小时收入/单次服务收入是我们关注的一项核心变量，一方面由于薪酬的特殊设计，使得其成为客服服务质量与效率的综合评价指标；另一方面，收入是最可感知，劳动者最为关心的核心指标。

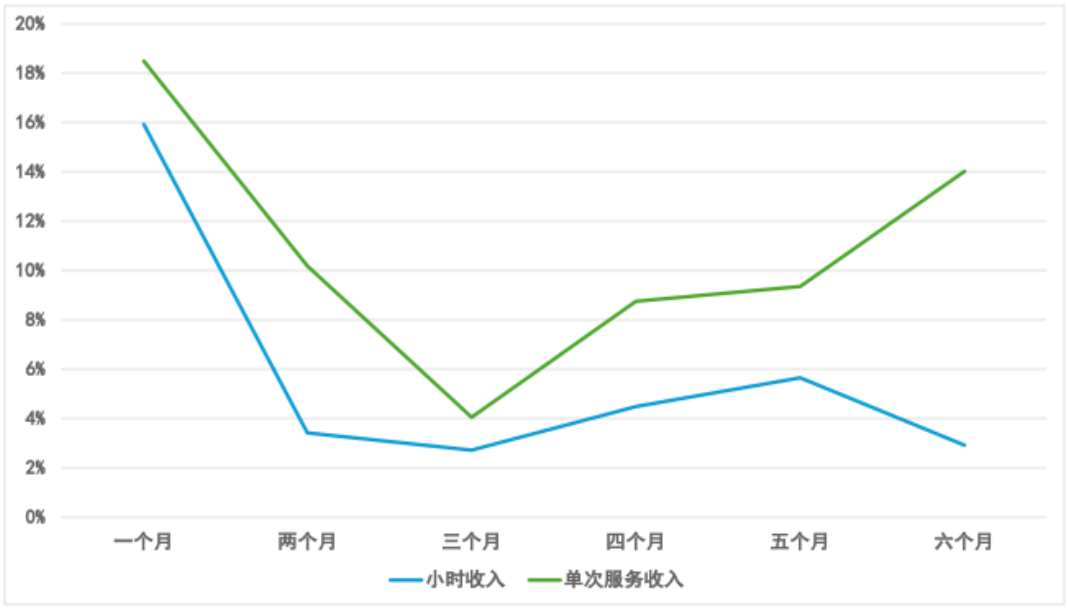

图 2 AI 训练对于客服收入的影响

上图展现了 AI 训练对于客服收入影响在上岗后六个月的变化情况，图中纵轴为 AI 训练相较于没有 AI 的反事实带来的收入增益。在各个时间尺度，以各类收入口径计量，AI 训练对于收入的影响均是正面的，以小时收入计，最大的正面增益发生在上岗后首月：15.93%。随着上岗时间的增加，AI 训练带来的增益在逐步收敛，但是单次服务收入的增益则不降反升，上岗后前六个月，AI 训练使得每次服务获得的平均薪酬上升了 14.02%。

2.对客户满意度的影响

客服工作的最终目标是解决客户问题，提升客户满意度，保证业务的良好运行。故而客服在工作中服务的客户的满意度，也是一个核心的度量指标。客户满意度的指标主要有两种：按键满意度，即在服务结束后，使用数字、星级、好差评按钮等方式对客服表现进行评价；调查满意度，即在服务结束后一段时间，通过应用内消息、短信等方式对客户进行调研，让其对服务进行打分。

不论按键满意度还是调查满意度，均会存在一个召回率的问题：并不是每次服务都会获得评价。故而我们选取的核心考核指标是，服务每小时/每个工作日获得的平均差评数，这基于一个朴素的逻辑：做出好评的客户并不一定真的觉得服务很好，但是做出差评的客户大概率是真的不开心；同时想做好评的人可能因为疏忽遗忘保护隐私等诸多考量不予好评，但是想做差评的人大多会突破重重阻碍反馈不满。

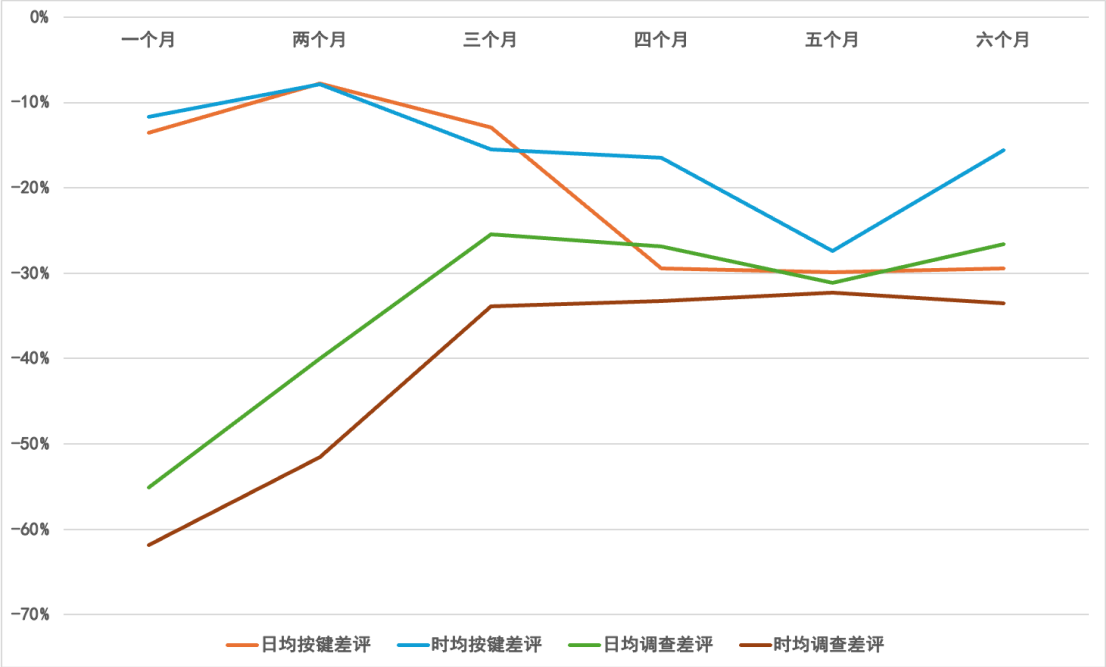

图 3 AI 训练对客户差评数的影响

各个维度而言，AI 训练均会显著降低用户的差评，在上岗后当月，AI 训练使得客服日均按键差评的数量降低了 13.56%；而在上岗后六个月内的日均按键差评的数量降低了 29.46%；而若以日均调研差评率计算，这两个数字分别为 55.12%与 26.57%。

较为有趣的一点是，调研差评率和按键差评率的 AI 影响在较长时间窗口内区别不大，但是在上岗初期有较大的差别，这一区别可能来自于调研与按键不同的反馈成本：按键反馈只是在界面上选择，而调研反馈则往往需要点击链接、填写问卷、提交等等复杂操作，复杂操作蕴含的不满是更大的。换言之，AI 训练对客户不满行为的抑制作用，在严重错误情境下可能更为显著。而这里的“严重错误”，则更可能是在服务态度等情感方面而非业务熟练等事实错误层面。这只是对于当前结果的一个合理猜想，更加严谨的研究有赖于后续对相关文本的语义分析。

3. 对服务标准程度的影响

客服在服务质量的另外一个维度，则是服务标准程度。由于客户在信息上处于劣势，其满意度更多地反映是其主观判断，而满意并不代表着客服反应的信息准确，服务流程标准、合规。一个形象的例子便是大模型中的幻觉，大模型幻觉

即是让客户满意但不能称之为标准的服务。对于客服服务的标准程度，在实践过程中，业务方安排专职专业人员对部分客服的服务记录进行人工校验。与差评类似，我们的主要考核指标是日均质检发现问题和时均质检发现问题。

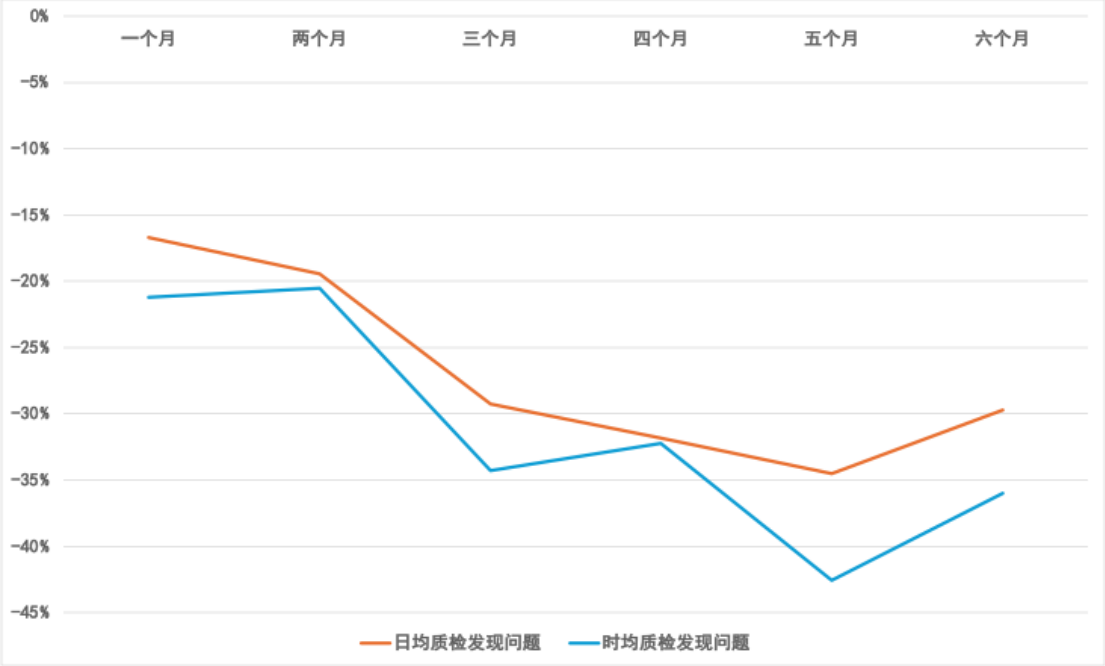

图 4 AI 训练对质检发现问题的影响

AI 训练使得，客服在上岗当月日均质检不合格数量下降 16.72%，这种优势在上岗后六个月的区间内持续，六个月整体不合格数量下降 29.70%。相较于服务评价，AI 训练对于服务标准程度的影响更加显著、稳健；同时 AI 训练对服务标准度的正面影响是随着时间增长放大的，这从一定程度上说明，更加灵活的 AI 剧本并没有使得服务标准化缺失，反倒因为其生动的方式形成了更加深刻、持久的影响。

（二）分人群的 AI 训练影响

近年来，数字鸿沟的影响引起了学者的广泛关注。数字鸿沟是指不同群体在信息通信技术（如互联网、智能终端等）的可获得性、使用能力及由此带来的经济社会收益上存在显著差距。故而 AI 训练是否带来了数字鸿沟，即在不同人群中产生不同的影响，值得特别关注。

我们采取的方法是，利用因果森林的特性，研究 AI 训练在不同人群中的条件平均处理效应（CATE）。因果森林在计算 CATE 的优势主要在于：它能够在非参数框架下灵活捕捉异质性处理效应，通过随机森林的集成机制降低因果树估计的

方差，从而在保持模型可解释性的同时提高估计的稳健性与精度；此外，它内置“忠诚样本分割”（honest splitting）方法，有效避免过拟合，确保对 CATE 的推断更具统计一致性。

我们主要关注三个方面的区分：性别、年龄、所在区域。

1.不同性别的 AI 训练影响差异

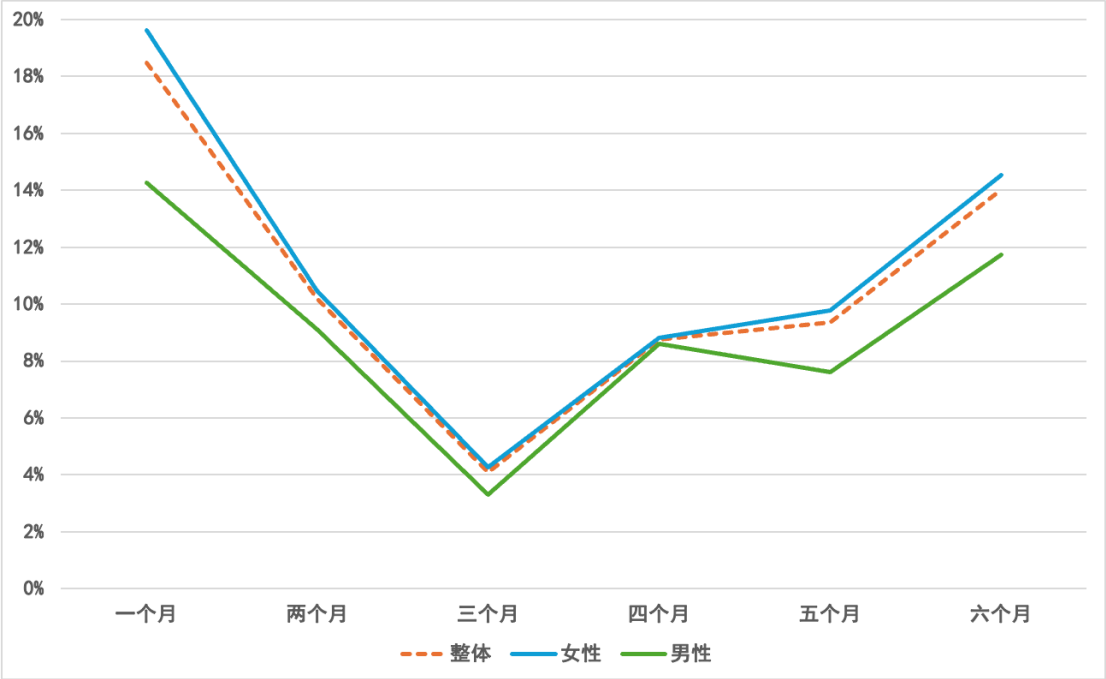

图 5 AI 训练对于收入影响的性别差异

上图展示了 AI 训练对于不同性别的客服，每次服务的平均收入的影响。结果表明，两性的劳动者收入均能得益于 AI 训练，但女性的增益效果更大。得益于 AI 训练，女性客服上岗后六个月内的件均收入增加了 14.52%，而男性群体这个数字为 11.75%。

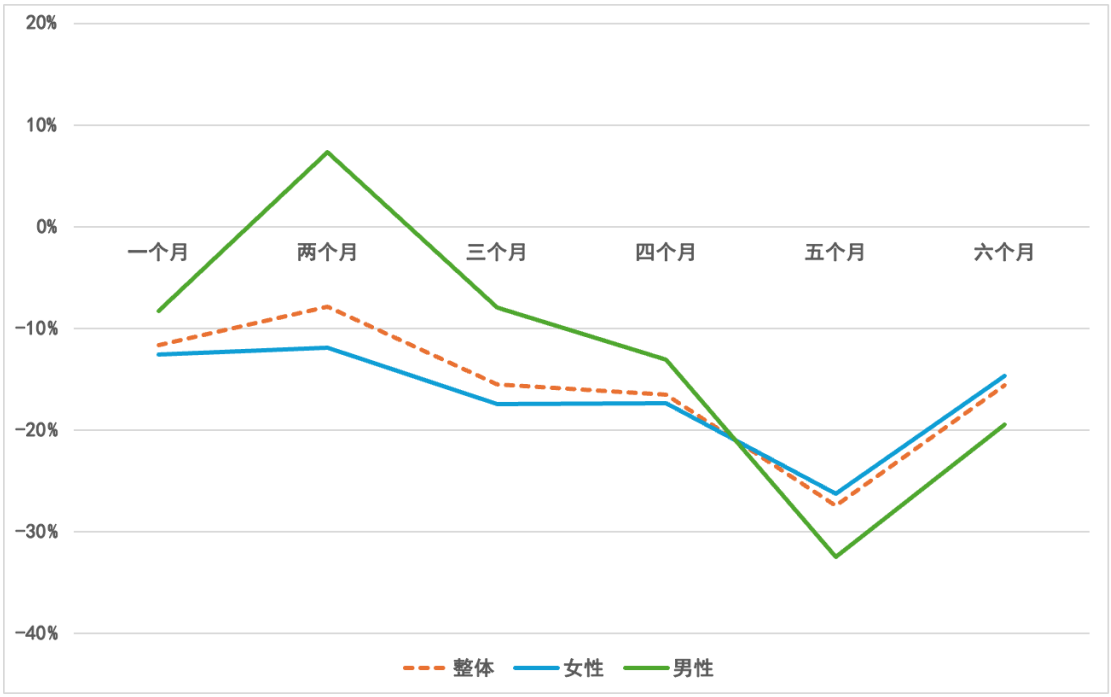

图 6 AI 训练对客户差评数量影响的性别差异

上图展示了 AI 训练对于客服平均每小时获得的差评数量的影响，在两性之间的差异，整体而言，男性则在服务满意度上更多的获益于 AI 训练，上岗后六个月时均差评量，降低了 19.4%。最有意思的一点是，在服务的前四个月，男性的获益是相对更小的，特别是前两个月的整体水平，AI 训练对于男性的影响是负面的（他们相较于不经历 AI 训练的反事实情况，收获了更多的差评）。同样现阶段只能提出一个猜测：这在某种程度上反映了一些逆反心理：“非不能也，不为也”——经历了一个月的工作，发现大可不必如训练时那般小心翼翼，但松懈带来了差评，差评又让人恢复了对于规矩的记忆。

2. 不同年龄段的 AI 训练影响差异

综合考虑现实意义、数据样本描述统计特征，我们将研究对象划分为 30 岁以下、30-45 岁、45 岁以上三个组别。

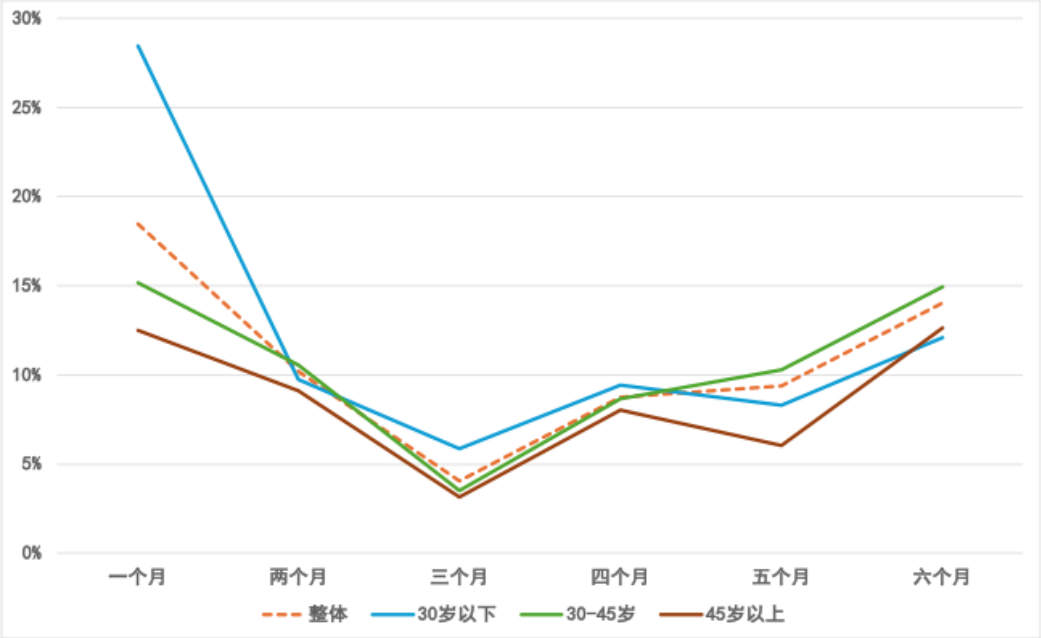

图 7 AI 训练对收入影响的年龄差异

上图展示了 AI 训练对于不同年龄阶段的客服，每次服务的平均收入的影响。数字鸿沟并没有在 AI 训练的场景中出现，除上岗当月外，不同年龄段的平均收入受 AI 训练的影响几乎没有区别。45 岁以上，岗后六个月内的每次服务的平均收入增加了 12.62%，30 岁以下群体、30-45 岁群体的这个数字则为 12.11%和 14.92%。

这种数据鸿沟的消弭应是预期中的，因为基于大模型的生成式 AI 的一个显著特征是使用门槛的大幅度降低，驱动智能的不再是代码而是自然语言，在 ChatGPT 3.5 的时代还要有特别的指令（Prompt）技巧来发挥模型的最大效率；之后愈发先进的模型的对于特殊指令的依赖愈发降低，只需要像对一个人那样细致、耐心、明确地描述需求便可以获得高质量答案；而到了智能体阶段，其工作流中已制定好的指令则可以进一步降低使用门槛——客户智能体的目标是模仿客服工作中可能遇到的客户，即像一个真人那样交流，那么使用这个智能体也只需要和“与人沟通”相一致的技巧，而这一技巧本不应存在数字鸿沟——那些年龄稍长的人往往有更强的沟通经验与处事技巧。

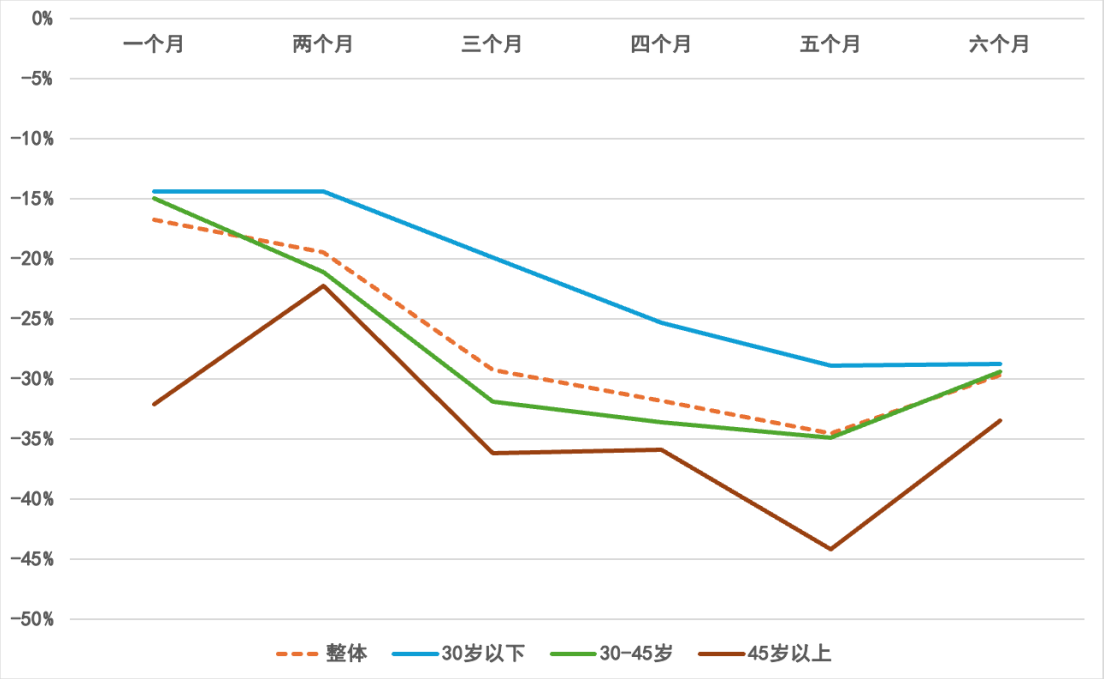

图 8 AI 训练对质检发现问题的影响的年龄差异

上图展示了 AI 训练对于服务标准性的影响在不同年龄段的差异，AI 的普惠性质则体现的更加明显，45 岁以上人群的服务标准性获得最大程度的改善，他们上岗后六个月日均质检不合格数量，下降了 33.46%，这一程度大于其他年龄组。这无疑是 AI 魅力的一大展现：受限于记忆力、技术理解力的差异，年长者在服务标准程度上原本处于劣势，但是却在 AI 训练中获益最多。这初步体现了 AI 对于能力培养的一种潜在优势：减弱了知识性的门槛，强化了与人沟通等通用能力的重要性，这不只有利于技能落后于时代的劳动者重返职场，也将助力劳动者在不同细分专业岗位上转换，找到与自己最匹配的职业。

3. 不同所在区域的 AI 训练影响差异

云客服的一项核心优势是，劳动者可以自由地选择工作地点，云客服往往可以选择居住在小城市、县城以降低生活成本，同时享受全国统一的薪资标准。云客服的广泛分布使得我们可以考察 AI 训练影响在城市等级、城乡间的差异。

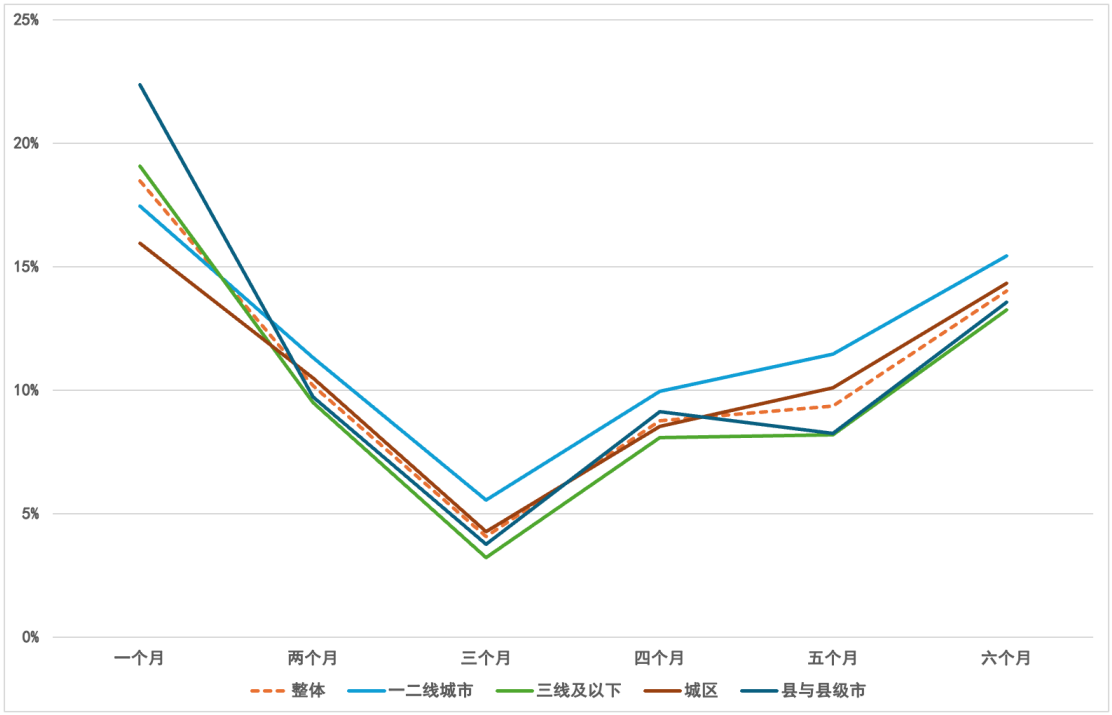

图 9 AI 训练对收入影响在地区间的差异

上图展示了 AI 训练对于客服平均每次服务收入的影响，在地区间的差异。我们主要考察了两个维度上的区分：按照《第一财经周刊》评定的城市分级，我们将客服分为两组：处于一线、新一线、二线城市；处于三线及以下城市。同时又将客服按所在区县的六位行政区划代码，分为两组：处于城区；处于县（旗）、县级市。研究发现，地理上的区分几乎未带来任何显著的影响。其他收入维度的指标、客户满意度、服务标准程度的指标，均为展现出地理区域间的差别，受限于篇幅，此处省略。

（三）AI 训练作用机制的初步分析

AI 训练的一项核心优势在于可以基于精巧的设计，模拟的客户具备多样化的身份、性格与情绪，提升客服在高压情况下的表现。在研究样本期内，我们发现了 23 种身份：如白领、时尚达人、技术控、普通身份等。在本研究中，我们将其二分为普通身份与其他复杂身份。

同时，AI 客户被按照 DISC 分型[①]设计成四种人格。它将人的行为特征划分为四种基本类型：D（Dominance，支配型）、I（Influence，影响型）、S（Steadiness，稳健型）和 C（Conscientiousness，谨慎型）。支配型的人目标导向强，决策果

① 美国心理学家威廉·莫尔顿·马斯顿博士在 1928 年提出，Marston W M. Emotions of normal people [M]. New York, NY: Harcourt Brace & Company, 1928.

断，善于把握全局；影响型的人外向乐观，善于沟通，倾向于通过关系和情感影响他人；稳健型的人注重合作与和谐，耐心可靠，强调团队与稳定环境；谨慎型的人重视规则与细节，逻辑性强，追求精确与高标准。在本研究中，我们将人格二分为稳健型与其他复杂性格。

同样，在剧本设计中，AI 客户涉及多种情绪转换，这些情绪包含许多种类型：乐观、暴躁、低落……等等，我们识别出其中两个集合：负面情绪与对抗性情绪，对抗性是负面的一个子集；同样，一个剧本可能涉及到多阶段的情绪，比如客户会从平静转变到低落继而到暴躁最后归于和解，在我们观察的样本中，一个剧本最多包含了七个情绪阶段。如果一个剧本包含 3 个或以上情绪阶段，我们将其定义为多阶段情绪剧本。

对于每一个 AI 训练的客服，我们构造一个哑变量（dummy）来衡量其是否得到了特别 AI 的训练：特殊 AI 训练时长占比超过 AI 其所经历 AI 时长训练的一半以上。

1. AI 身份与性格设计的影响

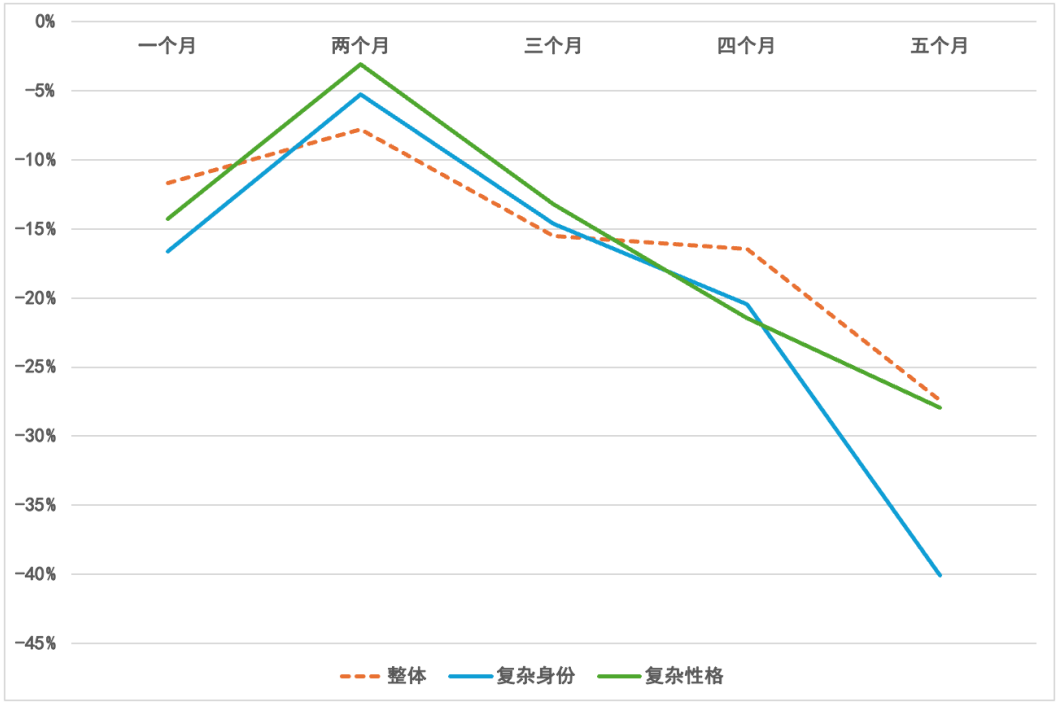

图 10 AI 身份与性格对于差评数量的影响

上图展示了 AI 训练对于客服平均每小时获得的差评数量的影响，在不同身份、性格上差异。结果显示，复杂性格和复杂身份的设计显著改善了客服在工作

中的客户评价。得到复杂身份剧本训练的客服五个月差评数量下降了 40.08%，复杂性格则带来了 27.91%的下降，均强于 AI 训练平均的 27.42%的效果。

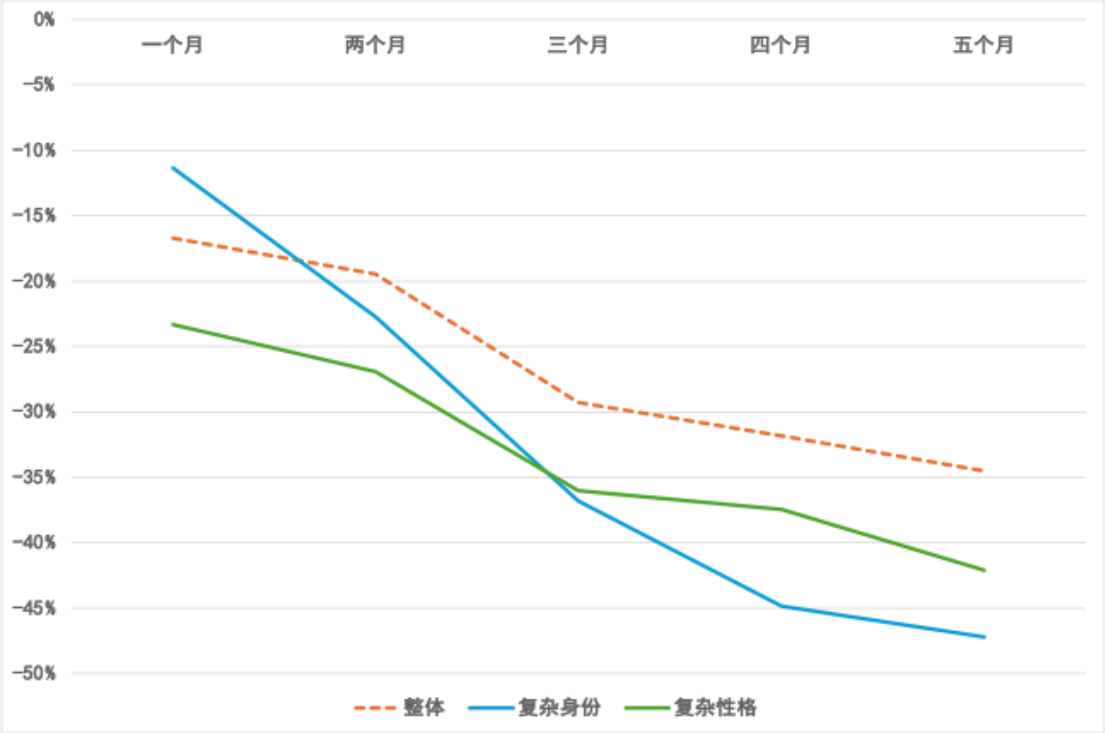

图 11 AI 身份与性格对于质检不合格数量的影响

上图展示了 AI 训练对于客服服务标准性的影响，在不同身份、性格上差异。得到复杂身份剧本训练的客服五个月质检不合格数量下降了 47.22%，复杂性格则带来了 42.08%的下降，均强于 AI 平均的 34.51%的效果。

2.AI 情绪设定的影响

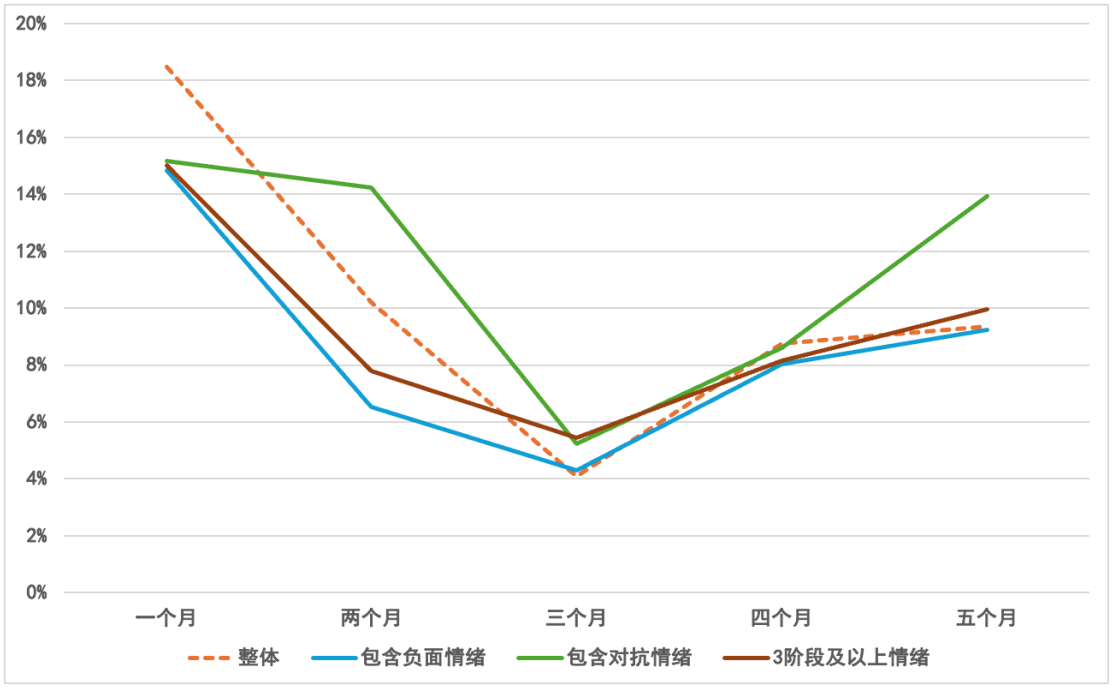

图 12 AI 情绪对于收入的影响

上图展现了 AI 情绪对于客服收入的影响。对抗性情绪的 AI 训练显著地改善了客服的收入，得到对抗情绪剧本训练的客服，五个月件均收入提升了 13.92%，大幅强于 AI 整体的 9.37%。而负面情绪、多阶段情绪的训练则提升不大。

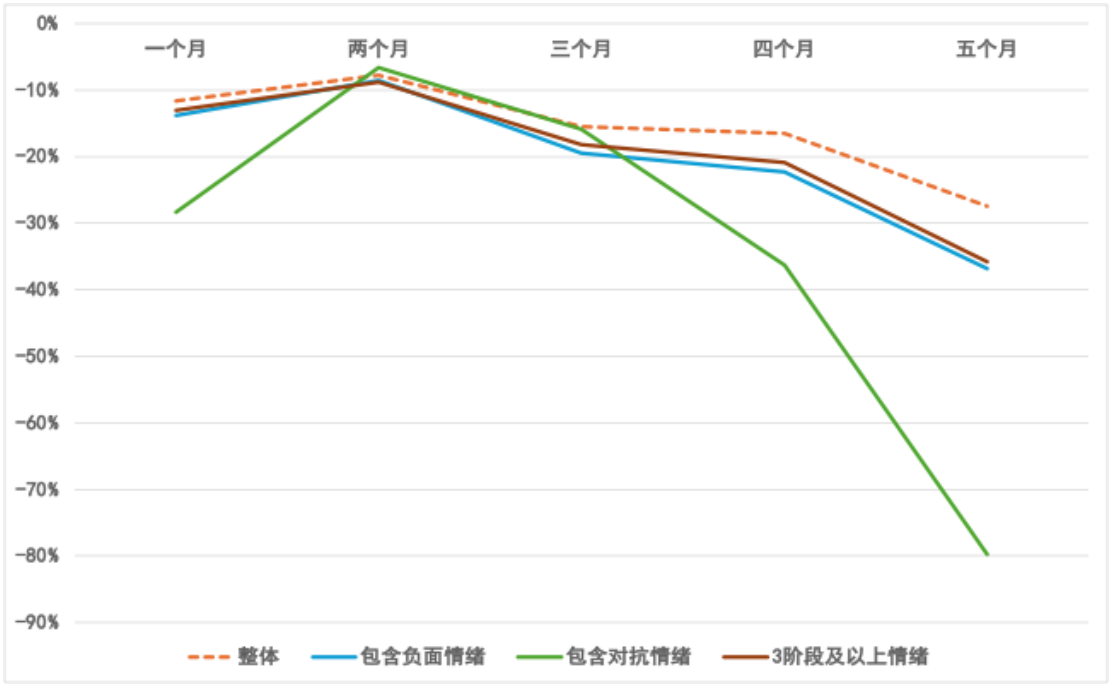

图 13 AI 情绪对于差评数量的影响

对抗性情绪的 AI 训练显著地改善了客服的客户评价，得到对抗情绪剧本训练的客服，五个月客户差评数量下降了 79.79%，负面情绪与多阶段情绪带来约 36%的差评减少，均强于 AI 训练整体带来的 27.42%的改善。对抗性情绪对于差评的减少是预期内，而减弱的幅度是超预期的。同时减少趋势随时间的变化则更为有趣，随着时间的累计，对抗情绪 AI 训练对于区间内平均差评数量的抑制作用逐步加强，从一定程度上佐证了，AI 训练是在能力上达到了“治本”级别的改进，而非对于工作流程、结果的“治标”。

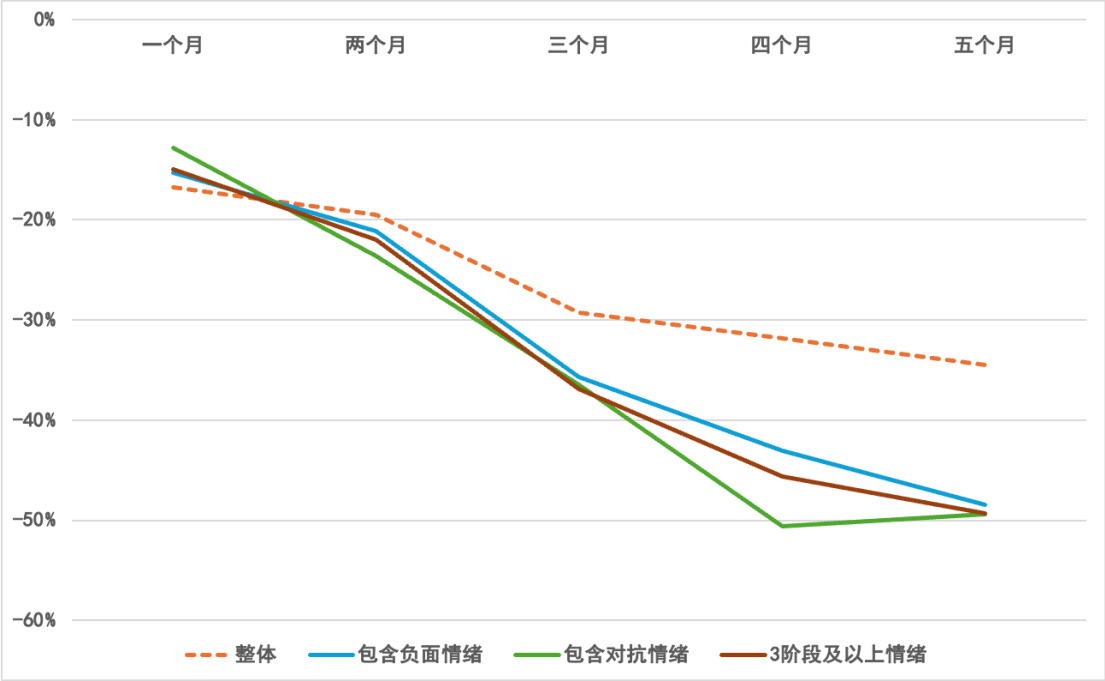

图 14 AI 情绪对于质检发现问题的影响

上图显示，AI 情绪的设计对于服务标准性的提升则更为一致、稳健，得到对抗情绪、负面情绪、多阶段情绪剧本训练的客服，五个月质检不合格数量下降了 49.40%、48.44%、49.33%，均强于 AI 训练整体带来的 34.51%的改善。

五、总结与讨论

至此，我们已经从整体、分人群、AI 复杂设计等多个角度揭示了 AI 训练对于客服能力及工作表现的提升作用。研究结果表明，AI 训练能够有效地提升新入职客服在上岗后的收入，提升其服务客户的满意度，增进其服务的标准性，那些更充分发挥生成式人工智能优势、贴近现实精巧设计的 AI 训练取得了更好的效果；同时，针对性别、年龄、所在区域的差异分析表明，AI 训练不存在数字鸿沟，甚至呈现出对弱势群体提升更大的普惠性质，这对我们如何利用相关技术，

提升劳动者技能，促进高质量就业有所启发。

技术发展对于就业冲击的问题是无法回避的，但技术引起的问题应首先尝试用技术去解决。生成式人工智能技术对使用门槛的降低是前几次人工智能浪潮未曾经历的，应当把握转型期的机遇窗口，充分利用数字经济过去十数年发展累积的数据沉淀与基础设施，着力开发针对劳动者技能提升、转换的 AI 训练应用、流程、体系，以智能手段，应对智能发展过程中的问题。

同时，应倡导在发展人工智能时，依照领域现实情况，设定多元价值取向。比如在事关大量人员就业岗位的行业，应当强调“就业优先”的理念。但这一追求并不是违背企业追求利润的“与虎谋皮”或者恐惧技术进步的“因噎废食”，而是规避市场参与者因片面追寻利益而导致市场失灵的“宏观调控”。这一论述来自以下一个有趣的发现：

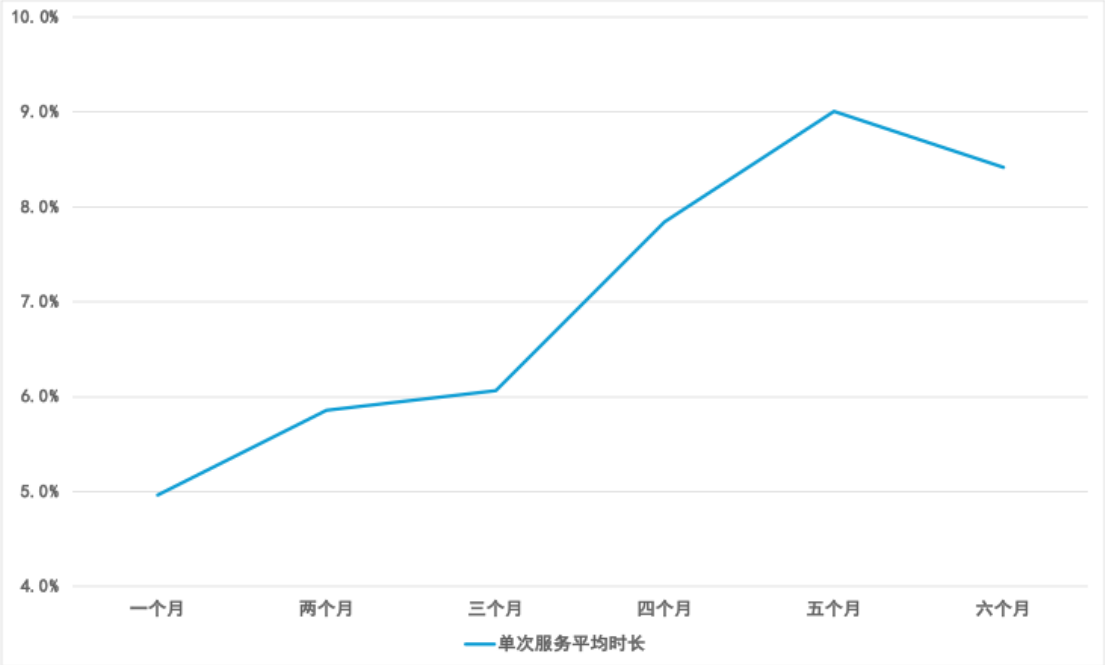

图 15 AI 训练对单次服务平均时长的影响

上图展示了 AI 训练对于单次服务平均时长的影响，整体而言，上岗后六个月，AI 训练使得服务时长增长了 8.42%。这一点是有些反直觉的，以至于我们需要把它特意放作最后一张图展开讨论。在一般场景下，单次客服时间较长往往被当作负面信号：需要更长时间，才能够“解决一个问题”，这无疑加重了企业提供客服的成本，降低了“效益”。然而上一节各个角度的分析结果表明，AI 训练的效果都是积极，为什么这里相左？

我们不妨再思考一下企业的“效益”与客户获得的“问题解决”，企业管理对于一般读者可能陌生，但是找客服去解决问题应当都有所经历，在相当多的时候并不只是在寻求一个疑问的解答，而是同时或者更看重一个情绪的平复。在训练台本中，多次出现一个“流动性”的关键词，经询问业务专家得知，这是其内部对于一类场景的特殊描述：用户的诉求无法解决，比如因为信用问题无法获得消费信贷的额度。而这时可能尽快完成一个标准答案的交付，并不能够说是“问题解决”，也无从提升企业的“效益”。

故而我们提出一个猜测，AI 训练带来的单次服务时长的提升，并非是客服熟练度不足的体现，而是他们在服务过程中包含了更多情绪共鸣、倾听理解的人文关怀，这一点的最终证实需要后续研究分析对话记录，但本次报告的发现是可以在一定程度上支持这一思路的。这样的取舍并非来自于价值观的训导，而是在于薪酬机制的精心设计——我们的研究结果表明，客服在完成效率上的损失，带来的是单次服务收入、单位时间收入的提升。

就此我们更进一步地将讨论扩展到 AI 对齐目标之上，现阶段再强大的智能也离不开人类意志的对齐，即 AI 的行动目标是满足人类为之预设的偏好，在一些领域更高更快更强天然地意味着更好的结果，而在一些满足人类需求的行业，AI 应用的目标是否该被锚定在效率之上？不同主体、不同时间窗口的思考会得到不同的结果，但人类经济发展史无数次地证明了单纯依靠市场调节的局限性，那么 AI 领域应也无法例外，对于 AI 对齐目标的“宏观调控”，应当被重视起来：不同企业、行业的 AI 在追求自身利益、恪守基本道德底线的同时，应当为社会整体的福利提升，对齐何种目标，值得讨论。

参考文献：

[1] Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo. “Artificial Intelligence and Jobs: Evidence from Online Vacancies.” Journal of Labor Economics, vol. 40, no. S1, 2022, pp. 293 - 340.

[2] Athey, Susan, and Guido W. Imbens. “Recursive Partitioning for Heterogeneous Causal Effects.” Proceedings of the National Academy of Sciences, vol. 113, no. 27, 2016, pp. 7353 - 7360.

[3] Babina, Tania, et al. "Artificial intelligence, firm growth, and product innovation." Journal of Financial Economics 151 (2024): 103745.

[4] Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. “GPTs Are GPTs: Labor Market Impact Potential of LLMs.” Science, vol. 384, no. 6702, 2023, pp. 1306–08.

[5] Felten, Edward, Manav Raj, and Robert Seamans. “A Method to Link Advances in Artificial Intelligence to Occupational Abilities.” American Economic Association Papers and Proceedings, vol. 108, 2018, pp. 54–57.

[6] Felten, Ed, Manav Raj, and Robert Seamans. "How will language modelers like ChatGPT affect occupations and industries?." arXiv preprint arXiv:2303.01157 (2023).

[7] Korinek, Anton, and Donghyun Suh. Scenarios for the Transition to AGI. NBER Working Paper no. 32255, National Bureau of Economic Research, 2024.

[8] Wager, Stefan, and Susan Athey. “Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests.” Journal of the American Statistical Association, vol. 113, no. 523, 2018, pp. 1228–1242.

[9] 张丹丹，于航，李力行，胡佳胤，莫怡青，and 李泓孛. 《中国人工智能技术暴露度的测算及其对劳动需求的影响——基于大语言模型的新证据》. 管理世界，no. 7，2025，pp. 59–70.