北京大学中国经济研究中心
China Center for Economic Research
讨论稿系列
Working Paper Series

C2026007 2026-08-17

薪酬结构对大语言模型采纳 及劳动生产率的影响

韩旭 易君健

摘要

本文研究薪酬结构如何影响劳动者对大语言模型的采纳，以及大语言模型接入带来的 生产率增益如何随薪酬结构而变化。本文开展了一项 2 × 3 实验室随机实验，同时改变大语言模型接入条件与薪酬结构。参与者被随机分配至有或无大语言模型接入的条件，并分别面对固定支付、绩效支付或锦标赛激励。本文得到三个主要发现。第一，相较于固定支付，绩效支付和锦标赛激励显著提高了劳动者对大语言模型的采纳。第二，大语言模型接入在三类薪酬结构下均显著提高了劳动生产率。第三，大语言模型接入带来的生产率增益在绩效关联更强的薪酬结构下更大，表明大语言模型的作用取决于劳动者所处的激励环境。分组分析进一步表明，绩效支付和锦标赛激励通过不同渠道发挥作用：前者主要在劳动者认为大语言模型能够提高绝对绩效时促进其使用，后者则主要在大语言模型有望形成相对绩效优势时发挥作用。总而言之，本文强调了一个直观但重要的思想：大语言模型的生产率后果不仅由技术本身决定，也由组织工作的经济制度决定。

关键词：薪酬结构；大语言模型；劳动生产率

JEL 分类号：C91; J24; J33

薪酬结构对大语言模型采纳及劳动生产率的影响*

韩旭† 易君健‡

2026 年 8 月

*易君健感谢国家自然科学基金（项目批准号：72595871, 72533001）的资助。文中观点及可能存在的错误均由作者负责。

†韩旭（Xu Han），斯坦福大学（Stanford University）。电子邮箱：xuhanx@stanford.edu。

‡易君健（Junjian Yi），北京大学中国经济研究中心，北京大学国家发展研究院，北京大学全球健康发展研究院。电子邮箱：junjian.yi@gmail.com。通讯作者。通讯地址：北京市海淀区蔚秀园路北京大学国家发展研究院（承泽园）。

薪酬结构对大语言模型采纳及劳动生产率的影响

摘要 本文研究薪酬结构如何影响劳动者对大语言模型的采纳，以及大语言模型接入带来的生产率增益如何随薪酬结构而变化。本文开展了一项 2×3 实验室随机实验，同时改变大语言模型接入条件与薪酬结构。参与者被随机分配至有或无大语言模型接入的条件，并分别面对固定支付、绩效支付或锦标赛激励。本文得到三个主要发现。第一，相较于固定支付，绩效支付和锦标赛激励显著提高了劳动者对大语言模型的采纳。第二，大语言模型接入在三类薪酬结构下均显著提高了劳动生产率。第三，大语言模型接入带来的生产率增益在绩效关联更强的薪酬结构下更大，表明大语言模型的作用取决于劳动者所处的激励环境。分组分析进一步表明，绩效支付和锦标赛激励通过不同渠道发挥作用：前者主要在劳动者认为大语言模型能够提高绝对绩效时促进其使用，后者则主要在大语言模型有望形成相对绩效优势时发挥作用。总而言之，本文强调了一个直观但重要的思想：大语言模型的生产率后果不仅由技术本身决定，也由组织工作的经济制度决定。

关键词：薪酬结构；大语言模型；劳动生产率
JEL **分类号**：C91; J24; J33

The Effects of Compensation Structures on Large Language Model Adoption and Labor Productivity

Abstract: This paper studies how compensation structures affect workers' adoption of a large language model (LLM) and how the productivity gains from LLM access vary across compensation structures. We conduct a laboratory randomized experiment with a 2×3 design that varies both LLM access and compensation structure. Participants are randomly assigned to either an LLM-access or no-LLM-access condition and to one of three compensation structures: lump-sum payment, performance-based payment, or tournament incentives. We document three main findings. First, relative to lump-sum payment, performance-based payment and tournament incentives significantly increase LLM adoption. Second, LLM access substantially improves labor productivity across all three compensation structures. Third, the productivity gains from LLM access are larger under compensation structures that place greater weight on performance, indicating that the effects of LLM access depend on the incentive environment in which workers operate. Subgroup analyses further suggest that performance-based payment and tournament incentives operate through different channels. Performance-based payment primarily encourages LLM use when workers perceive the LLM as improving their absolute performance, whereas tournament incentives matter more when LLM use is expected to create a relative performance advantage. Overall, this paper highlights a simple but important idea: the productivity effects of large language models depend not only on the technology itself, but also on the economic institutions that govern work.

Keywords: Compensation structures; Large language models; Labor productivity
JEL Codes: C91; J24; J33

一、 引言

大语言模型正在迅速进入各类工作场景，并成为影响劳动过程和组织生产方式的重要新技术。越来越多的研究表明，大语言模型能够在知识密集型任务中显著改善劳动者表现，例如提高文本写作质量、缩短任务完成时间，或为经验较少的劳动者提供实时辅助 (Noy and Zhang, 2023; Bick et al., 2025; Brynjolfsson et al., 2025; Chatterji et al., 2025)。然而，大语言模型带来的生产率提升并非均匀发生。现有研究揭示了不同劳动者、任务类型和工作环境之间显著的效应异质性，表明大语言模型的经济后果不仅取决于技术本身，也取决于劳动者在何种条件下选择采纳该技术。由此产生了一个尚未被充分回答的重要问题：当劳动者获得大语言模型接入机会时，何种因素决定其是否选择使用大语言模型？进一步地，这种采纳选择又在什么条件下能够转化为更大的劳动生产率提升？

薪酬结构是理解上述问题的一个自然切入点。在多数工作场景中，大语言模型使用既不是自动发生的，也并非完全没有成本。劳动者需要决定是否采纳大语言模型、在多大程度上使用大语言模型，以及如何将模型输出与自身判断结合起来。这些决策很可能取决于提高绩效所带来的边际回报。在绩效边际回报较低的环境下，劳动者可能缺乏足够动机去承担使用大语言模型所伴随的货币或认知成本；而在绩效关联更强的环境下，同一项工具可能变得更有吸引力。更重要的是，不同薪酬机制还可能以不同方式塑造大语言模型使用。例如，绩效支付奖励产出的绝对改善，而锦标赛激励奖励相对表现。因此，大语言模型带来的生产率后果可能不仅取决于劳动者是否获得技术接入，也取决于劳动者在使用技术时面对何种薪酬结构。

本文研究薪酬结构如何影响劳动者对大语言模型的采纳，以及大语言模型接入所带来的生产率增益如何随薪酬结构而变化。本文的核心思路是，薪酬结构改变劳动者提高绩效的边际回报，从而改变其将大语言模型作为生产率提升工具的采纳激励。一旦大语言模型的使用成为劳动者的内生行为选择，大语言模型接入带来的生产率增益就不应被视为某种一成不变的技术效应。本文通过一项同时随机改变大语言模型接入和薪酬结构的实验来研究这一问题，从而将大语言模型采纳视为劳动者对薪酬结构的内生行为反应，而非技术可得性的机械结果。

本文与两支文献密切相关。第一，本文与关于大语言模型及其劳动市场影响的新兴文献相关。近期研究记录了大语言模型异常迅速的扩散及其潜在生产率效应。利用美国全国代表性调查数据，Bick et al. (2025) 表明大语言模型已经在劳动年龄人口中快速普及；Chatterji et al. (2025) 利用大规模对话数据进一步发现，大语言模型在工作相关场景中的使用大量集中于写作、信息获取和实用建议等活动。从更一般的技术与任务视角看，这与劳动经济学关于技术改变任务结构的经典框架相一致：计算机技术更容易替代可编码的例行任务，并与非例行的问题解决和沟通任务形成互补 (Autor et al., 2003)；任务型技术变迁模型也强调，技术既可能自动化既有任务，也可能创造新的任务，其总体劳动市场后果取决于受到影响的任务范围以及新任务如何组织 (Acemoglu and Restrepo, 2018, 2020)。因此，大语言模型的总体经济影响不仅取决于其技术能力，也取决于实际采纳的广度和强度，以及微观层面的生产率收益能否推广至更加复杂、情境依赖更强的任务 (Acemoglu, 2025)。

已有微观研究进一步表明，大语言模型接入能够显著影响劳动生产率。Brynjolfsson et al. (2025) 研究客服人员使用大语言模型助手后的表现，发现大语言模型辅助能够显著提高劳动者产出，并且对经验较少或初始技能较低的劳动者帮助更大。Noy and Zhang (2023) 在专业写作任务中发现，ChatGPT 能够显著缩短任务完成时间并提高产出质量。与此同时，大语言模型的有效性也具有明显的任务边界：当任务与模型能力较好匹配时，大语言模型可能显著提高生产率；对于依赖深层背景、默会知识或复杂判断的任务，其作用则可能减弱甚至产生负面影响 (Dell'Acqua et al., 2023)。软件开发领域的随机试验同样发现了显著的平均生产率增益以及不同劳动者和工作流程之间的异质性 (Cui et al., 2026)。此外，即使在大语言模型暴露程度较高的职业中，实际采纳仍然并不均衡 (Humlum and Vestergaard, 2025)。这些研究清楚地表明大语言模型可能影响劳动生产率，但对于制度环境和薪酬结构如何影响劳动者是否选择采纳大语言模型，仍缺乏系统证据。特别是，已有研究中的“大语言模型接入”效应往往同时包含劳动者是否使用工具、如何使用工具以及如何重新配置努力和注意力等多个行为边际；即使接入状态由外生方式决定，实际使用仍然可能是内生的。

第二，本文与人事经济学（personnel economics）中关于薪酬合同和劳动者行为的经典文献相关。人事经济学长期强调，薪酬合同会改变劳动者努力、选择、注意力配置和生产率 (Gibbons, 1998; Prendergast, 1999)。在多任务委托–代理框架中，基于不完全绩效指标的激励还可能改变劳动者在不同活动之间的努力配置 (Holmstrom and Milgrom, 1991)。大量经验研究进一步表明，薪酬结构会实质性地影响劳动者表现。Lazear (2000) 发现，由固定工资转向绩效工资能够通过提高努力和改变劳动者选择两条渠道提高生产率；实地实验同样发现，计件工资相对于固定

工资能够显著提高劳动者产出 (Shearer, 2004)。锦标赛理论则将基于排名的薪酬机制刻画为取决于相对绩效而非绝对产出的激励方式 (Lazear and Rosen, 1981)。Bandiera、Barankay 和 Rasul 的一系列企业实地实验进一步表明，改变劳动者和管理者面对的薪酬结构不仅会影响平均生产率，还会改变社会互动、产出分布、管理者注意力配置以及团队结果 (Bandiera et al., 2005, 2007, 2010, 2013)。更广泛地说，人事经济学和组织经济学研究强调，技术与组织实践之间具有重要的互补关系，新技术所带来的生产率收益往往取决于相应的工作组织和人力资源实践 (Bresnahan et al., 2002; Lazear and Shaw, 2007)。然而，这支文献尚未系统研究薪酬合同如何影响劳动者对大语言模型这一新型生产工具的采纳。

本文将上述两支此前相对独立发展的文献联系起来。我们开展了一项包含 289 名参与者的实验室随机实验，采用 2×3 实验设计。参与者被随机分配至大语言模型接入组或无接入组，并同时被随机分配至三种薪酬结构之一：固定支付、绩效支付或锦标赛激励。参与者随后依次完成三项用于刻画不同类型知识工作的任务：文稿校对与修改、视频播放量预测和预算分配。在大语言模型接入组中，我们观察参与者是否选择使用实验平台内置的大语言模型工具，并记录其与模型的交互行为；在全样本中，本文估计不同薪酬结构下大语言模型接入对劳动生产率的影响。

本文得到三个主要发现。第一，绩效关联更强的薪酬结构提高大语言模型采纳。相较于固定支付，在能够使用大语言模型的参与者中，绩效支付和锦标赛激励均提高了大语言模型使用率。第二，大语言模型接入显著提高劳动生产率。在绩效得分、质量调整后的绩效得分和产出数量等多个结果变量上，大语言模型接入均带来了显著的正向影响。第三，大语言模型接入带来的生产率增益并不独立于薪酬结构。相较固定支付，大语言模型接入在绩效支付和锦标赛激励下带来的生产率提升更大。

本文进一步利用参与者信念和不同任务之间的差异开展机制分析。参与者对开放式问题的回答提供了提示性证据，表明绩效支付和锦标赛激励可能通过不同渠道影响大语言模型采纳。绩效支付主要在劳动者认为大语言模型能够提高其绝对任务表现时促进大语言模型使用。相比之下，锦标赛激励的作用更具策略性：当劳动者预期大语言模型能够帮助其形成相对绩效优势时，锦标赛更能够促进大语言模型使用；而当劳动者担心大语言模型会使不同竞争者的答案趋于相似时，这一作用会减弱。上述解释与近期关于薪酬结构影响算法建议依赖的实验研究相一致 (Greiner et al., 2026)。Greiner et al. (2026) 发现，相较固定支付，绩效支付和锦标赛合同会提高被试对预先给定的算法建议的依赖。相对于该研究，本文有两点推进。第一，本文考察大语言模型在多类知识工作任务中的使用，从而覆盖不同生产领域中的完整任务过程。第二，本文关注参与者是否主动向大语言模型寻求帮助这一采纳决策，而非对预先给定算法建议的被动依赖，并将这一采纳选择直接与实际劳动生产率结果相联系。

本文的贡献主要体现在三个方面。第一，现有大语言模型研究主要关注向劳动者提供大语言模型接入机会所产生的平均处理效应，而本文提供了薪酬结构如何影响大语言模型采纳的因果证据，将大语言模型使用理解为劳动者对制度环境作出的内生行为反应，而非技术接入后的机械结果。第二，本文表明，大语言模型的生产率效应取决于劳动者所处的薪酬结构，从而将组织激励设计与新技术的实际生产率回报联系起来。第三，本文表明，不同薪酬合同可能诱发不同的大语言模型使用方式：一些薪酬结构可能促使劳动者更广泛地将大语言模型作为提高任务表现的生产工具，而另一些薪酬结构则可能诱导更加选择性和策略性的大语言模型使用。更一般地说，本文结果表明，组织不能将大语言模型视为一种独立于制度环境的技术投入；大语言模型的实际影响还取决于工作场所中的薪酬结构。

本文的实验设计还具有两项测量优势。其一，三项任务覆盖文字编辑、预测判断和管理决策等不同类型的知识工作，并尽可能依据预先设定的评分规则、真实结果或明确约束评价产出，因而能够同时考察绩效、质量和数量，而不依赖单一、难以核验的主观评价。其二，实验在线下实验室完成，参与者使用统一设备和内置工具。相较于难以观察实验外工具使用的远程环境，这一设计更有利于控制未授权的大语言模型使用，并准确区分“获得接入”与“实际采纳”两个行为环节。

本文结论的外部推广应考虑到如下实验场景的限制。实验对象主要为大学生，任务为受控环境下的短时认知任务；大语言模型采纳设置了较小但显著的进入成本，产出则依据明确评分规则评价。因此，本文结果对于与上述特征相近的知识型工作环境最具解释力，不宜直接推广至所有工作场景。

本文余下部分安排如下：第二部分介绍实验设计；第三部分介绍数据与变量；第四部分报告大语言模型采纳和劳动生产率的主要结果；第五部分开展机制分析；第六部分总结全文并讨论研究局限与展望。

二、 实验设计

为考察薪酬结构如何影响劳动者的大语言模型采纳及其劳动生产率后果，本文开展了一项实验室随机实验。实验采用 2 × 3 的实验设计，同时操纵参与者是否能够使用大语言模型，以及其在完成任务时面对的酬金发放规则。具体而言，参与者被随机分配到大语言模型接入组或无大语言模型接入组，并进一步被随机分配到固定支付、绩效支付或锦标赛激励三种酬金规则之一。实验要求参与者依次完成三项知识型工作任务。下文将依次介绍样本对象、随机分组、实验处理、实验任务、评分规则和实验流程。

（一） 样本对象与随机分组

本研究的目标人群是那些在实际工作中可能独立使用或与大语言模型协作完成任务，并具备基本大语言模型交互能力的劳动者。此类工作通常包括写作、信息处理、分析和决策支持等认知型、非例行任务。它们相较于高度依赖体力劳动或线下互动的工作，更容易受到大语言模型的影响 (Bick et al., 2025; Chatterji et al., 2025)。在实际招募中，本文主要从北京若干高校招募参与者。虽然大学生群体不能完全代表现实劳动市场中的所有劳动者，但其作为未来劳动市场参与者，较可能进入受大语言模型影响较大的知识型岗位，也通常具备一定的大语言模型使用经验 (Baek et al., 2024)。因此，对于研究大语言模型在知识型工作中的采纳和生产率效应，大学生样本具有较好的适配性。

本实验包含两个随机分组维度。第一个维度是大语言模型接入。参与者被随机分配到可使用内置大语言模型助手的实验组，或无法使用大语言模型助手的对照组。第二个维度是薪酬结构。本文将薪酬结构具体化为酬金发放规则，因为酬金合同决定了劳动者提高绩效的预期边际收益，从而影响其努力程度以及是否采纳生产率提升工具 (Lazear and Rosen, 1981; Holmstrom and Milgrom, 1991; Lazear, 2000; Bandiera et al., 2007, 2013)。参与者被随机分配到固定支付、绩效支付或锦标赛激励三种酬金规则之一。如图1所示，上述设计形成六个实验组：有大语言模型接入下的固定支付组、绩效支付组和锦标赛组，以及无大语言模型接入下的固定支付组、绩效支付组和锦标赛组。

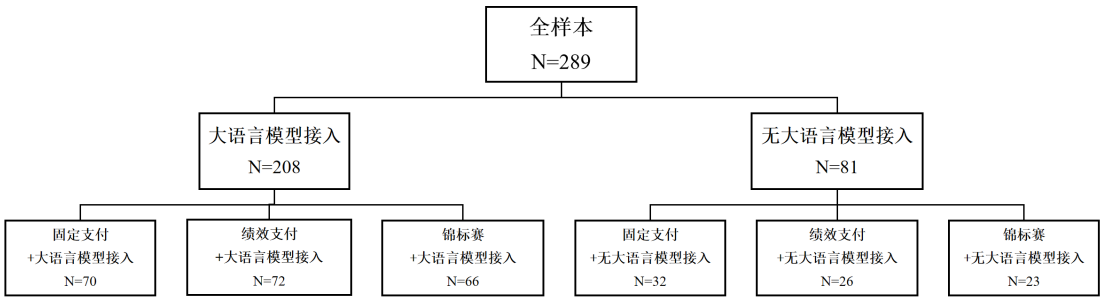


图 1 实验设计示意图

注：本图展示实验设计。参与者被随机分配到大语言模型接入组或无接入组，并被随机分配到固定支付、绩效支付或锦标赛三种酬金规则之一，从而形成 2 × 3 实验设计。

每日第一场实验开始前，在当日报名参与者中由计算机程序进行个体随机分组。各实验日的分组概率在实验前预先设定，并在整个实验期间保持不变，因此该程序不会系统性影响不同实验组的样本构成。为获得更丰富的大语言模型采纳和人机交互数据，本文将大语言模型接入组的分配概率设置得相对更高：70% 的参与者被分配到大语言模型接入组，30% 被分配到无大语言模型接入组。在给定大语言模型接入状态后，参与者被均匀分配到三种酬金规则下。若参与者缺席或提交无效答案，其原有分组不重新抽取。因此，本文的识别依赖于缺席和无效作答在各实验组之间随机发生的假设。

（二） 实验处理

1. 大语言模型接入

被分配到大语言模型接入组的参与者，可以在完成每项任务时与实验平台内置的大语言模型助手进行交互。该助手由 DeepSeek-V3.2 模型驱动，该模型为实验期间可用的最新 DeepSeek 模型。为避免模型品牌认知对参与者使用行为产生影响，本文将大语言模型工具直接嵌入实验平台，并未向参与者披露模型名称；大语言模型助手在交互过程中也不会主动透露其模型身份。

相应地，被分配到无大语言模型接入组的参与者无法使用大语言模型辅助模块，需要在没有大语言模型帮助的情况下完成相同任务。附录A展示了大语言模型接入组和无大语言模型接入组的实验界面。除了是否能够使用大语言模型工具（以及与大语言模型接入直接相关的界面和使用说明）之外，两组参与者面对的实验环境、任务材料和实验流程均保持一致。在整个实验过程中，对于无大语言模型接入组，实验说明中尽可能减少了对大语言模型的提及。因此，本实验随机分配的大语言模型接入状态的差异与实验环境中的其他因素相互独立。

需要说明的是，大语言模型接入虽然由实验随机分配，但实际大语言模型使用是参与者的内生行为选择。参与者可以在每项任务中自行决定是否与大语言模型助手交互。实验平台记录参与者在每项任务中是否发生大语言模型交互，以及完整的人机交互日志。为避免大语言模型使用率过高，并更好地模拟现实工作场景中使用大语言模型可能伴随货币成本、认知成本或社会成本的情况，本文对每项任务的大语言模型使用设置 0.25 元的进入成本。该成本每项任务最多收取一次，即同一任务内后续多轮交互不会产生额外费用。平台界面会明确提示这一成本，并在参与者每项任务首次尝试使用大语言模型时再次提醒。

2. 酬金规则

为操纵薪酬结构，参与者被随机分配到三种酬金规则之一：固定支付、绩效支付和锦标赛激励。这三种规则覆盖了从绩效关联较弱到较强的制度变化，有助于比较不同薪酬机制是否会诱导不同程度的大语言模型采纳，以及大语言模型接入在不同薪酬机制下是否具有不同的生产率效应。每项任务开始前，实验平台都会向参与者展示其所面对的具体酬金规则。

第一，固定支付（lump-sum payment）。固定支付组旨在刻画边际激励较弱的环境。在这一规则下，参与者只要任务表现达到最低通过标准，即可获得固定酬金。该最低标准被设置为较容易达到的水平。因此，在达到该标准之后，进一步提高任务表现不会带来额外酬金，参与者面对的边际激励较弱。[1]

第二，绩效支付（performance-based payment）。绩效支付组将酬金与任务表现更紧密地联系起来。在这一规则下，参与者的酬金随其测度绩效提高而增加，任务表现到酬金的映射关系单调递增。与固定支付相比，绩效支付为参与者提高任务表现提供了更强的边际激励。[2]

第三，锦标赛（Tournament）。锦标赛组的酬金取决于参与者的相对表现，而非仅取决于其绝对表现。实验结束后，锦标赛组参与者被随机分入十人小组，并根据其在小组内的任务表现排名决定酬金。参与者在完成任务时并不知道同组其他参与者的身份和表现。排名越高，获得的酬金越高。与固定支付和绩效支付相比，锦标赛激励引入了相对竞争因素，因为参与者不仅需要考虑自身产出，还需要考虑其他参与者的预期表现。

（三） 实验任务

所有参与者均需完成三项知识型工作任务。本文选择任务时主要基于以下考虑：第一，每项任务均对应现实工作中可能出现的认知型劳动，并来自不同的工作场景；第二，任务产出能够依据预先设定的评分规则进行评价；第三，大语言模型在三项任务中均可能提供帮助，但仅依靠大语言模型并不足以稳定产生与高能力劳动者相当的高质量答案。

具体而言，任务 1 是新闻稿修改与编辑，任务 2 是视频播放量预测，任务 3 是预算分配及解释。三项任务分别对应描述性、分析性和管理性认知工作，并且在脑力判断与机械执行的相对重要性、以及与大语言模型能力的匹配程度上存在明显差异。这种任务间差异有助于降低合并分析对单一任务特征的依赖，也为考察薪酬结构对大语言模型采纳及其生产率效应的任务异质性提供了空间。

三项任务在大语言模型辅助的适配程度上也存在差异。任务 1 与当前大语言模型的语言处理能力最为契合。任务 2 相对不适合大语言模型辅助，因为视频播放量预测依赖对受众行为和内容热度的判断，而当前大语言模型可能缺乏此类任务所需的具体背景知识。任务 3 介

[1]相关文献中类似规则有时被称为 fixed payment (Shearer, 2004; Bandiera et al., 2007)。现实劳动市场中的固定工资仍可能包含解雇风险、晋升激励或职业声誉等隐性激励，但这些因素在本文的一次性实验场景中基本不存在。因此，本文设置较低的最低通过标准，以在保证参与者基本投入的同时维持较弱的边际激励。

[2]在相关文献中，类似规则也常被称为计件工资（piece-rate payment）(Lazear, 2000; Shearer, 2004)。由于本文任务表现并非由单一可计数产出衡量，而是包含多个维度的评分，因此本文采用更宽泛的“绩效支付”这一表述。

于两者之间：大语言模型能够帮助组织思路并生成连贯的文字说明，但高质量答案仍需要参与者对可行性、约束条件和优先次序作出判断。

下面具体介绍三个任务的细节。

第一，新闻稿修改与编辑。任务 1 要求参与者将一篇存在问题的学校新闻稿修改为可发布稿件。参与者需要在 10 分钟内完成该任务。任务内容包括错误识别、事实一致性核查、语句改写和整体编辑润色，因此较好地模拟了白领工作中的文字编辑任务，也适合用于考察大语言模型如何影响纠错准确性和写作质量。为捕捉薪酬机制在节省时间方面的作用，任务 1 的酬金规则还在保证质量的条件下奖励更快完成任务。

该新闻稿中预先嵌入了若干错误，因此任务 1 的绩效得分主要取决于参与者成功识别并修改错误的比例。这些错误包括语法错误、重复表述等大语言模型通常较擅长发现和修改的问题，也包括与若干经典文学作品相关、需要一般背景知识辅助判断的问题。后者可能更容易被人类参与者识别。因此，该任务为人类判断与大语言模型辅助之间的互补性留下了空间。

第二，视频播放量预测。任务 2 是一项多子题预测任务。参与者会看到若干组在线视频标题及其基本信息，例如视频时长和发布时间，并需要预测在 YouTube 或 Bilibili 等平台上哪一个视频更可能获得更高播放量。该任务共有 12 道小题，限时 5 分钟。每个小题中，参与者需要在 A、B 两个选项中作出选择，并简要说明其预测理由。

在该任务中，产出数量可以自然地用完成的小题数量衡量。视频标题和基本信息改编自 Bilibili 平台上的真实视频，并去除了所有私人或可识别信息。由于每个视频的实际播放量可以观察到，本文能够以真实播放结果作为客观标准，对参与者的预测表现进行评价。

第三，预算分配。任务 3 要求参与者代表一个公益基金会，在四项面向中学生的心理健康干预项目之间分配固定预算，并提供预算分配方案及文字解释。该任务限时 10 分钟，需要参与者进行结构化推理、优先级排序和书面解释。相较于任务 1，任务 3 更少依赖事实纠错，而更多依赖判断、权衡和政策性推理。

任务材料中明确指出，出于捐赠方和宣传需求，项目需要在三周内产生可观测效果，例如能够收集到学生反馈。在四个备选方案中，有一个方案在给定预算和时间约束下几乎无法实施，另一个方案则相对而言不够合理。因此，分配给这两个方案的资金会按照不同程度进行折减。任务 3 的绩效得分根据折减后的有效资金总额计算。由于部分方案涉及中学、大学生招募和在线视频等内容，参与者在判断时可以调用自身经验，这也为参与者自身判断与大语言模型辅助之间的互补性提供了空间。

（四） 评分规则和结果变量

本文关注两类结果变量：大语言模型采纳和劳动生产率。大语言模型采纳，也即大语言模型使用，是大语言模型交互在扩展边际上的重要指标，具体定义为在可使用大语言模型的情况下，参与者是否在某项任务中与大语言模型助手发生交互。劳动生产率依据各任务预先设定的评分规则进行衡量，包括反映客观任务表现的绩效得分、在绩效得分基础上纳入答案质量奖惩的质量调整后绩效得分，以及反映产出规模的产出数量指标。[3]

如前文所述，不同任务的评分规则根据其具体特征分别设定。在任务 1 中，绩效主要取决于参与者识别并修改新闻稿中预设错误的能力，质量奖惩规则还会考察整体写作质量。在任务 2 中，绩效依据各小题预测准确率计算，产出数量则由完成的小题数衡量。在任务 3 中，绩效取决于预算分配决策和文字解释的质量，其中相对不合理的选项对应的资金会根据特定规则进行折减，最终得分基于折减后的有效资金总额计算。

上述评分设计有两方面目的。第一，它使每项任务都能够按照自身生产逻辑得到评价。第二，它使本文能够在任务之间构造可比较、较全面的质量和数量指标，并在后续实证分析中使用这些指标衡量劳动生产率。

（五） 实验流程

图2概括了本文实验流程。参与者首先在线报名并填写一份简短问卷，内容包括基本人口统计特征。每个实验日的报名在当天第一场实验开始前截止。报名成功后，参与者来到实验室并进入实验说明阶段。该阶段包括实验说明、知情同意、酬金发放规则的介绍。大语言模型接入组的参与者还会与大语言模型助手进行热身交互。实验说明阶段还包含一道理解检验题，用于确认参与者理解其所面对的酬金规则。

[3] 上述生产率指标评分主要用于本文的实证分析。用于计算实验酬金时，本文对评分进行了单调变换，使固定支付组的最低通过标准对大多数参与者而言较容易达到。该变换在实验前预先设定，不改变参与者面对的激励方向。对于固定支付组，参与者在实验说明中被告知该标准是一个较容易达到的最低通过标准。

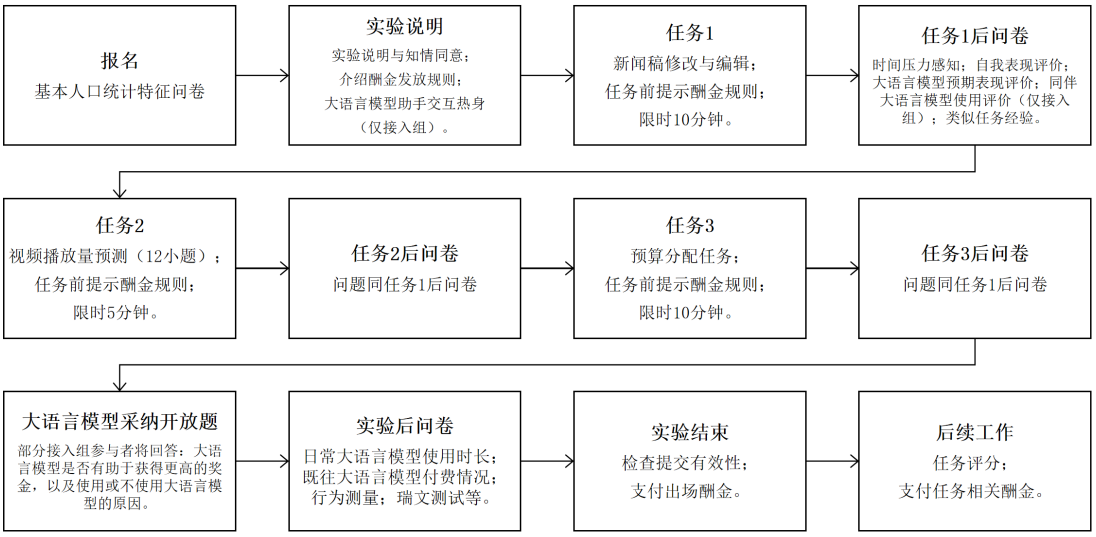


图 2 实验流程示意图

注：本图展示实验流程。参与者先完成报名和实验说明，随后依次完成三项主要任务及任务后问卷。同伴大语言模型使用信念等问题仅向接入组询问；部分接入组参与者还回答关于采纳决策的开放式问题。

随后，参与者按照固定顺序依次完成三项限时任务。[4] 每项任务开始前，平台都会再次提醒参与者具体酬金规则。每项任务结束后，参与者需要回答一组简短的任务后问卷。问卷内容包括其感知到的时间压力、对自身任务表现的评价、对大语言模型预期表现的评价，以及是否曾接触过类似任务。对于大语言模型接入组，任务后问卷还会询问参与者对其他参与者大语言模型使用情况的信念。上述问题被设计得较为简短，以便在尽量不干扰实验流程的情况下，趁参与者对任务体验仍较新鲜时收集其信念信息。这些变量使本文能够比较具有不同大语言模型绩效预期和同伴使用预期的参与者在大语言模型采纳和劳动生产率方面的差异。

三项主要任务结束后，部分大语言模型接入组参与者需要回答一道关于大语言模型采纳的开放式问题，解释其是否认为大语言模型能够帮助自己获得更高酬金，以及为什么选择使用或不使用大语言模型。最后，参与者填写实验后问卷，报告既往大语言模型使用经验以及若干用于后续实证分析的基线特征。完成全部流程后，实验结束并结算酬金。

三、 数据与变量

本部分介绍本文实证分析所使用的样本、变量构造和描述性统计结果，并检验随机分组后的样本平衡性。

（一） 样本

本文实验于 2026 年在北京大学的实验室线下开展，共持续 9 个实验日，每个实验日包含 2 至 4 场实验，最终得到 289 名有效参与者样本。[5] 其中，208 名参与者被分配到大语言模型接入组，81 名参与者被分配到无大语言模型接入组。由于每名参与者均依次完成三项任务，本文主要分析在参与者–任务层面展开，共包含 867 个观测值。大语言模型采纳行为仅能在大语言模型接入组中观察到，因此关于大语言模型采纳的分析基于有大语言模型接入的子样本，共包含 624 个参与者–任务层面观测值。

参与者主要来自北京大学及北京市其他高校，样本以本科生和低年级研究生为主。考虑到本文实验任务具有较强的文字处理、信息分析和决策支持特征，该样本与本文所关注的知识型劳动场景具有较高相关性。同时，实验室环境保证了参与者面对相同的实验流程、界面环

[4]固定任务顺序可能带来学习效应、疲劳效应，或使参与者在后续任务中更加熟悉实验界面和大语言模型助手。作为稳健性检验，本文主要结果在仅使用任务 1 样本时仍基本一致，而任务 1 不受此前任务经验影响。对于合并分析而言，所有参与者面对相同任务顺序，处理状态随机分配，且回归中控制任务固定效应，因此该问题可以在一定程度上得到缓解。不过，合并估计结果仍应理解为固定任务顺序设计下的处理效应。

[5]共有 310 名参与者来到实验室参加实验。其中，21 名参与者因中途退出或提交无效答案而被排除在最终样本之外。上述样本损耗在不同实验组之间大体平衡。

境和任务材料，有助于减少现实工作场景中其他混杂因素的影响。

本文数据具有参与者–任务层面的面板结构。实验处理在参与者层面随机分配，且同一参与者将在数据中出现三次。因此，在后续合并回归中，本文将标准误聚类到参与者层面。与此同时，本文在合并任务的回归模型中控制任务固定效应，以吸收不同任务在难度、时间限制和产出形式上的系统性差异。

（二） 变量

本文的结果变量主要分为两类。第一类是大语言模型采纳变量。在大语言模型采纳回归中，因变量为一个虚拟变量，若参与者在某项任务中与大语言模型助手发生交互，则该变量取值为 1，否则取值为 0。由于无大语言模型接入组无法使用大语言模型，该变量仅在大语言模型接入组中定义。

第二类是劳动生产率变量。本文主要使用三项标准化的任务层面结果变量衡量劳动生产率：标准化绩效得分、标准化质量调整后绩效得分和标准化产出数量。

第一，绩效得分。绩效得分根据各任务预先设定的客观评分规则构造。虽然不同任务的具体评分方式不同，但原始得分均被统一到 0 至 10 分量表上。任务 1 的绩效得分主要取决于参与者识别并修改新闻稿中预设错误的比例；任务 2 的绩效得分取决于视频播放量预测的准确性；任务 3 的绩效得分则基于预算分配中折减后的有效资金总额。

第二，质量调整后绩效得分。该变量在绩效得分基础上，进一步纳入基于扩展版评分规则的质量奖励和惩罚。相较于绩效得分，该指标更强调答案质量中的主观维度，例如论证的说服力、分析深度和文字表达的连贯性。

第三，产出数量。产出数量衡量参与者在任务中实际完成的产出规模。对于任务 1 和任务 3，产出数量以字数衡量；对于任务 2，产出数量以完成的小题数量衡量。

为便于在不同任务之间进行比较，本文将上述三项生产率指标在任务内标准化。因此，在后续回归中，回归系数可以直接解释为以标准差为单位的变化。

本文的主要解释变量为实验处理变量，包括大语言模型接入状态、酬金规则，以及二者之间的交互项。基线控制变量包括参与者年龄、性别、是否来自实验室所处的大学、日常大语言模型交互时长、是否曾为大语言模型服务付费、是否接触过类似任务、风险厌恶、损失厌恶，以及两道瑞文测试题的答题表现。其中，风险厌恶和损失厌恶通过两道彩票选择题度量：若参与者在相应问题中选择更确定的选项，则被定义为风险厌恶或损失厌恶。瑞文测试题用于捕捉参与者的基线认知能力，其中较简单和较困难两题分别构造为答对与否的虚拟变量。

（三） 描述性统计

表1报告了本文样本的主要描述性统计结果。Panel A 报告大语言模型采纳与劳动生产率相关变量，Panel B 报告参与者的处理前基线特征，Panels C–E 分别报告任务 1 至任务 3 的相关统计量。

首先，大语言模型使用并非普遍发生。在大语言模型接入组中，参与者在任务层面使用大语言模型的平均概率为 53.5%。这一事实表明，即使大语言模型使用的显性成本很低，参与者也并未将大语言模型视为自动默认选项，而是会根据任务环境和自身判断选择是否使用大语言模型。

其次，大语言模型采纳在不同任务之间存在明显差异。任务 1 中大语言模型使用率为 60.6%，任务 2 中为 30.8%，任务 3 中为 69.2%。这一差异与三项任务和大语言模型能力之间的匹配程度基本一致：新闻稿修改与预算分配更容易从语言生成、结构化表达和方案整理中受益，而视频播放量预测更依赖对具体内容热度和受众行为的判断。因此，较低的任务 2 大语言模型使用率也说明参与者能够根据任务特征调整大语言模型采纳行为。

再次，生产率结果变量在样本中呈现出较为充分的变异。全样本中，原始绩效得分均值为 5.504 分（满分 10 分），标准差为 2.406 分；质量调整后绩效得分均值为 4.714 分（满分 10 分），标准差为 2.351 分。

最后，参与者具有一定的大语言模型使用经验，但也存在显著异质性。参与者报告的日常大语言模型交互时长平均约为 1.58 小时，43.9% 的参与者曾为大语言模型服务付费。这说明样本并非完全不熟悉大语言模型的群体，但也并不只是由高度熟练的大语言模型用户构成。这种使用经验上的差异为本文考察薪酬结构是否会改变大语言模型采纳行为提供了有利条件。

表 1 描述性统计

变量	观测值	均值	中位数	标准差	最小值	最大值
Panel A. 大语言模型采纳与劳动生产率						
大语言模型使用（是 =1，仅限接入组）	624	0.535	1	0.499	0	1
绩效得分（0–10）	867	5.504	5.833	2.406	0	10
标准化绩效得分	867	0	0.150	1	-3.064	2.194
质量调整后绩效得分（0–10）	867	4.714	5	2.351	0	10
标准化质量调整后绩效得分	867	0	0.099	1	-2.737	2.443
标准化产出数量	867	0	-0.175	1	-2.113	7.258
完成时间（秒）	867	346.300	300	141.611	36	600
Panel B. 参与者特征						
年龄	867	21.284	21	2.950	18	34
女性（是 =1）	867	0.623	1	0.485	0	1
来自北京大学（是 =1）	867	0.630	1	0.483	0	1
报告的每日大语言模型交互时长（小时）	867	1.582	1	1.496	0.017	10
曾为大语言模型服务付费（是 =1）	867	0.439	0	0.497	0	1
风险厌恶（是 =1）	867	0.671	1	0.470	0	1
损失厌恶（是 =1）	867	0.588	1	0.492	0	1
瑞文测试 1（较易，答对 =1）	867	0.872	1	0.334	0	1
瑞文测试 2（较难，答对 =1）	867	0.398	0	0.490	0	1
Panel C. 任务 *1*：新闻稿修改与编辑						
大语言模型使用（是 =1，仅限接入组）	208	0.606	1	0.490	0	1
绩效得分（0–10）	289	5.281	5.556	2.631	0	10
标准化绩效得分	289	0	0.104	1	-2.007	1.794
质量调整后绩效得分（0–10）	289	4.634	4.722	2.442	0	9.5
标准化质量调整后绩效得分	289	0	0.036	1	-1.898	1.993
标准化产出数量	289	0	-0.547	1	-1.179	2.825
完成时间（秒）	289	382.799	379	159.247	40	600
产出字数	289	764.163	616	270.771	445	1,529
报告的时间压力（1–5）	289	2.789	3	1.106	1	5
对自身表现的评价（0–100）	289	78.394	80	12.749	30	100
对大语言模型表现的评价（0–100）	289	86.263	90	10.604	40	100
对同伴大语言模型使用率的判断（%）	208	66.240	70	28.117	5	100
曾接触类似任务（是 =1）	289	0.796	1	0.404	0	1

表 1 描述性统计（续）

变量	观测值	均值	中位数	标准差	最小值	最大值
Panel D. 任务 *2*：视频播放量预测						
大语言模型使用（是 =1，仅限接入组）	208	0.308	0	0.463	0	1
绩效得分（0–10）	289	4.694	5	2.038	0.833	9.167
标准化绩效得分	289	0	0.150	1	-1.894	2.194
质量调整后绩效得分（0–10）	289	3.749	3.333	2.218	0	9.167
标准化质量调整后绩效得分	289	0	-0.187	1	-1.690	2.443
标准化产出数量	289	0	0.378	1	-2.113	0.932
完成时间（秒）	289	281.606	300	43.742	50	300
产出字数	289	212.208	151	183.005	12	1,148
报告的时间压力（1–5）	289	4.349	5	0.961	1	5
对自身表现的评价（0–100）	289	63.540	65	20.126	8	100
对大语言模型表现的评价（0–100）	289	76.758	80	16.680	10	100
对同伴大语言模型使用率的判断（%）	208	58.760	60	28.295	5	100
曾接触类似任务（是 =1）	289	0.405	0	0.492	0	1
Panel E. 任务 *3*：预算分配						
大语言模型使用（是 =1，仅限接入组）	208	0.692	1	0.463	0	1
绩效得分（0–10）	289	6.536	7	2.133	0	10
标准化绩效得分	289	0	0.218	1	-3.064	1.624
质量调整后绩效得分（0–10）	289	5.760	5.950	1.922	0.5	10
标准化质量调整后绩效得分	289	0	0.099	1	-2.737	2.206
标准化产出数量	289	0	-0.153	1	-1.359	7.258
完成时间（秒）	289	374.460	402	163.395	36	600
产出字数	289	364.069	331	216.439	70	1,935
报告的时间压力（1–5）	289	2.401	2	1.009	1	5
对自身表现的评价（0–100）	289	77.574	80	13.231	30	100
对大语言模型表现的评价（0–100）	289	80.394	80	14.711	10	100
对同伴大语言模型使用率的判断（%）	208	70.635	80	23.655	10	100
曾接触类似任务（是 =1）	289	0.723	1.000	0.448	0	1

注：本表报告样本描述性统计。Panel A–B 报告全样本统计，Panel C–E 按任务分别报告。大语言模型使用变量仅在接入组中观察。绩效得分依据预先设定、以客观表现为核心的任务评分规则构造；质量调整后绩效得分在此基础上加入反映答案质量的奖励和惩罚，其中包含一定主观评价维度。产出数量在任务 1 和任务 3 中以字数衡量，在任务 2 中以完成的小题数量衡量。所有标准化变量均在任务内标准化。风险厌恶和损失厌恶由两道彩票选择题测量：若参与者选择更确定的选项，则相应指标取 1。瑞文测试 1（较易）和瑞文测试 2（较难）分别表示对应认知题是否答对。Panel C–E 中的同伴使用率信念仅向大语言模型接入组询问。

（四） 平衡性检验

为检验随机分组是否有效，本节进一步考察不同实验组在处理前特征上的平衡性。具体而言，本文以“固定支付且无大语言模型接入”组作为基准组，将年龄、性别、是否来自实验室所处的大学、日常大语言模型交互时长、是否曾为大语言模型服务付费、风险厌恶、损失厌恶以及两道瑞文测试题答题结果分别对各实验组虚拟变量进行回归，并在所有回归中控制实验场次固定效应。

表2报告了平衡性检验结果。总体来看，随机分组具有较好的平衡性。大多数估计系数相对于基准组均值较小，且在统计上不显著。在全部 45 个组间差异中，仅有 2 个在常规水平上显著，且不存在围绕某一实验组或某一基线变量的系统性不平衡。这说明本文实验随机分组总体有效。

由于处理状态由实验随机分配，组间均值差异本身即可识别处理效应。因此，基线控制变量并非识别所必需。不过，加入基线控制变量和实验场次固定效应有助于提高估计精度。为此，后文将同时报告简单回归结果，以及进一步加入基线控制变量和实验场次固定效应的估计结果。

四、 主要结果

本文的实证分析分为两步。第一步考察薪酬结构是否影响参与者的大语言模型采纳。由于该部分分析仅限于大语言模型接入组，所有参与者均可使用大语言模型，因此不同酬金规则下大语言模型使用率的差异可以解释为薪酬结构对大语言模型采纳决策的因果影响。第二步考察大语言模型接入对劳动生产率的影响。由于大语言模型接入状态由实验随机分配，在不同酬金规则下比较大语言模型接入组和无大语言模型接入组的任务表现，可以识别大语言模型接入在不同薪酬机制下的异质性生产率效应。

（一） 薪酬结构与大语言模型采纳

首先，本文考察在参与者能够使用大语言模型时，绩效关联更强的薪酬结构是否会提高其大语言模型采纳概率。图3报告了不同酬金规则下大语言模型使用率的原始均值比较。结果显示，相较于固定支付组，绩效支付组和锦标赛组的大语言模型使用率均明显更高。具体而言，相较于固定支付组，绩效支付组的大语言模型使用率约高 15 个百分点，锦标赛组约高 11 个百分点。其中，绩效支付组与固定支付组之间的差异在统计上显著，锦标赛组与固定支付组之间的差异则边际显著。这一模式初步表明，当酬金规则提高参与者从任务表现改进中获得的边际收益时，参与者更倾向于使用大语言模型。

为进一步控制基线特征和实验场次差异，本文估计如下线性概率模型：

$$\begin{aligned} LLMUse_{its} = \beta_1 PerformanceBased_i + \beta_2 Tournament_i \\ + \Gamma' X_{it} + \lambda_t + \delta_s + \varepsilon_{its}. \end{aligned} \tag{1}$$

式中，因变量表示参与者在某项任务中是否使用大语言模型；两个处理系数分别衡量绩效支付和锦标赛相对于固定支付的影响。模型控制参与者基线特征、任务固定效应和实验场次固定效应。由于同一参与者完成三项任务，标准误聚类到参与者层面。

表3报告了回归结果。第 (1) 和第 (2) 列使用三项任务的合并样本。结果显示，无论是否加入基线控制变量和实验场次固定效应，绩效支付均显著提高大语言模型采纳概率。在加入控制变量和实验场次固定效应后，绩效支付的系数为 0.153，且在 1% 水平上显著。相较于固定支付组 44.8% 的大语言模型使用率，该估计值意味着绩效支付使大语言模型使用率提高约三分之一。锦标赛激励的系数同样为正。在加入控制变量和实验场次固定效应后，锦标赛组的大语言模型使用率相较于固定支付组高 12.2 个百分点，这表明锦标赛激励也会提高大语言模型采纳概率。

表 2 随机分组平衡性检验

因变量:	年龄	女性（是 =1）	来自北京大学（是 =1）	每日大语言模型交互时长（小时）	曾为大语言模型服务付费（是 =1）	风险厌恶（是 =1）	损失厌恶（是 =1）	瑞文测试 1（答对 =1）	瑞文测试 2（答对 =1）
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
固定支付 + 大语言模型接入	-0.483	-0.008	-0.096	-0.768**	-0.080	-0.048	0.051	0.004	-0.013
	(0.586)	(0.112)	(0.100)	(0.351)	(0.117)	(0.109)	(0.113)	(0.078)	(0.113)
绩效支付 + 大语言模型接入	-0.347	0.017	-0.049	-0.363	-0.011	-0.095	0.190*	0.032	-0.047
	(0.573)	(0.109)	(0.098)	(0.342)	(0.114)	(0.107)	(0.110)	(0.076)	(0.110)
锦标赛 + 大语言模型接入	-0.080	0.058	-0.058	-0.421	-0.046	-0.074	0.115	0.063	-0.006
	(0.581)	(0.111)	(0.099)	(0.348)	(0.116)	(0.108)	(0.112)	(0.077)	(0.112)
绩效支付 + 无大语言模型接入	0.248	-0.186	-0.089	0.148	0.155	-0.034	0.154	0.066	0.053
	(0.681)	(0.130)	(0.116)	(0.407)	(0.136)	(0.127)	(0.131)	(0.090)	(0.131)
锦标赛 + 无大语言模型接入	-0.375	-0.049	-0.094	-0.245	0.152	0.056	0.002	-0.039	0.035
	(0.704)	(0.135)	(0.120)	(0.421)	(0.141)	(0.131)	(0.135)	(0.094)	(0.135)
固定支付 + 无大语言模型接入组均值	21.938	0.656	0.656	1.970	0.438	0.688	0.469	0.875	0.406
观测值	289	289	289	289	289	289	289	289	289

注：本表在参与者层面检验列标题所示处理前特征的平衡性。省略组为固定支付且无大语言模型接入组。风险厌恶、损失厌恶和瑞文测试变量的定义见表1。括号内为标准误，每个回归均控制实验场次固定效应。显著性水平：$^{*}\ p < 0.1$，$^{**}\ p < 0.05$，$^{***}\ p < 0.01$。

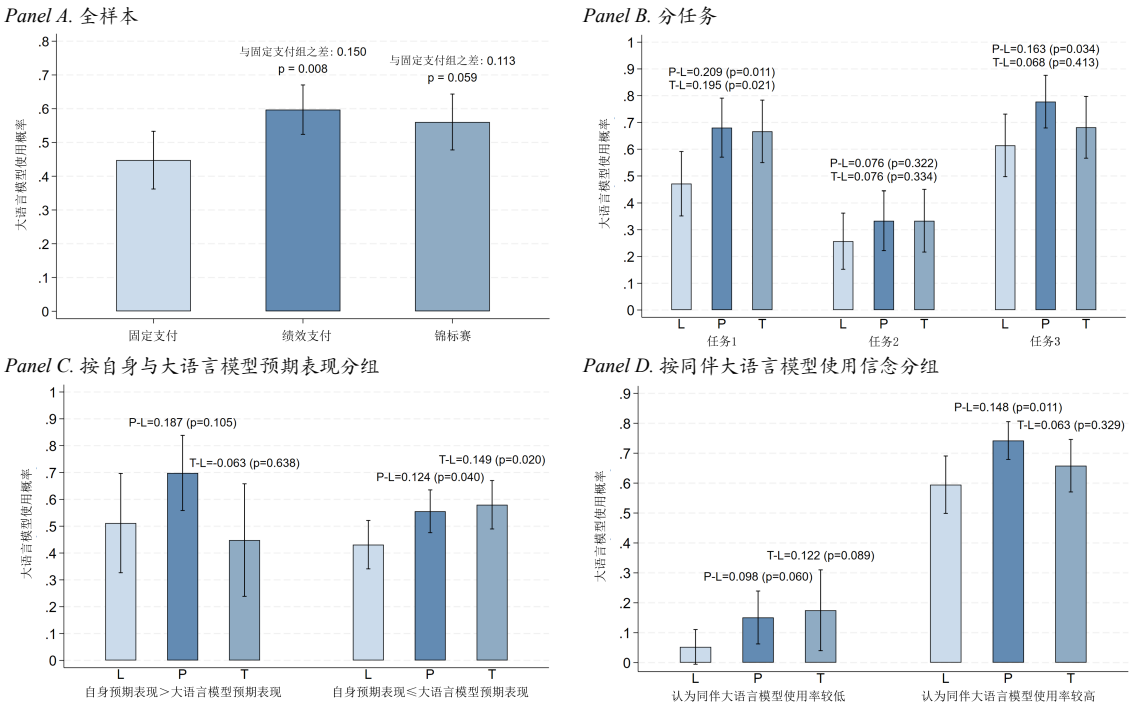


图 3 薪酬结构与大语言模型采纳：均值比较和分组分析

注：本图报告大语言模型接入组在不同薪酬结构下的平均使用概率。Panel A 为全样本比较；Panel B 按任务报告；Panel C 按参与者是否认为大语言模型预期表现高于自身分组；Panel D 按同伴使用信念分组。Panels B–D 中的 L、P 和 T 分别表示固定支付、绩效支付和锦标赛。误差线为 95% 置信区间；图中差异和 p 值均相对于固定支付组计算，标准误聚类到参与者层面。

表 3 薪酬结构与大语言模型采纳

因变量:	大语言模型使用（是 =1）							
	全部任务		任务 1		任务 2		任务 3	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
绩效支付	0.150***	0.153***	0.209**	0.231***	0.076	0.049	0.163**	0.174**
	(0.056)	(0.057)	(0.081)	(0.084)	(0.078)	(0.083)	(0.077)	(0.080)
锦标赛	0.113*	0.122**	0.195**	0.191**	0.076	0.078	0.068	0.088
	(0.059)	(0.057)	(0.083)	(0.086)	(0.080)	(0.085)	(0.079)	(0.081)
固定支付组均值	0.448	0.448	0.471	0.471	0.257	0.257	0.614	0.614
基线控制变量	否	是	否	是	否	是	否	是
实验场次固定效应	否	是	否	是	否	是	否	是
观测值	624	624	208	208	208	208	208	208
R^2	0.126	0.193	0.038	0.205	0.006	0.128	0.022	0.203

注：本表报告大语言模型接入组中薪酬结构对大语言模型使用的 OLS 回归结果。因变量为参与者是否在某项任务中使用大语言模型。固定支付组为省略组。基线控制变量包括年龄、性别、是否来自实验室所处的大学、每日大语言模型交互时长、是否曾为大语言模型服务付费、是否接触过类似任务、风险厌恶、损失厌恶及两道瑞文测试题的表现。第 (1)–(2) 列合并三项任务并控制任务固定效应；标准误聚类到参与者层面。显著性水平：$^{*}\ p < 0.1$，$^{**}\ p < 0.05$，$^{***}\ p < 0.01$。

第 (3) 至第 (8) 列进一步按任务分别估计。结果显示，薪酬结构对大语言模型采纳的影响在不同任务之间存在差异。在任务 1 中，绩效支付和锦标赛激励均显著提高大语言模型使用率；加入控制变量后，二者的系数分别为 0.231 和 0.191。在任务 2 中，两种绩效关联更强的薪酬结构的系数均为正，但数值较小且不显著。这可能是因为视频播放量预测任务与大语言模型能力的匹配程度较低，参与者认为大语言模型难以提供明显帮助。在任务 3 中，绩效支付显著提高大语言模型使用率，加入控制变量后的系数为 0.174；锦标赛激励的系数为正但不显著。总体而言，大语言模型采纳并非大语言模型接入后的机械结果，而是会随参与者所面对的薪酬结构和任务特征发生变化。

进一步地，本文考察薪酬结构对大语言模型采纳的影响是否随参与者的基线认知能力而变化。本文以两道难度不同的瑞文测试题近似衡量基线认知能力，其中瑞文测试 1 相对较易，瑞文测试 2 相对较难。图4分别按照参与者是否答对两道瑞文测试题报告薪酬结构对大语言模型采纳的异质性结果。

图4显示，薪酬结构的影响可能与参与者的基线认知表现存在非单调关系。Panel A 按较易的瑞文测试 1 分组：在未答对该题的参与者中，绩效支付和锦标赛对大语言模型采纳的影响均较弱；在答对该题的参与者中，两类薪酬结构对大语言模型使用率的促进作用更加明显。Panel B 按较难的瑞文测试 2 分组，呈现出不同的分组模式：在未答对该题的参与者中，绩效支付和锦标赛的影响相对较大，而在答对该题的参与者中，两类薪酬结构的影响均有所减弱。综合两个 Panel 的结果，薪酬结构对大语言模型采纳的影响似乎在基线认知表现居中的参与者中更为明显。

一种可能的解释是，不同认知能力的参与者对大语言模型使用价值的边际反应存在差异。认知能力较低的参与者可能较难判断或有效整合大语言模型提供的辅助，而认知能力较高的参与者可能已经形成较为稳定的任务策略，对大语言模型辅助的边际需求相对有限。相比之下，处于中间水平的参与者可能更容易随着使用大语言模型预期收益的变化而调整采纳决策。不过，这一解释仅具有提示性。本文的认知能力测量仅基于两道瑞文测试题，且分组后的子样本量较小，因此图4中的结果应被视为探索性的异质性证据。

需要强调的是，实验中的酬金规则并没有直接奖励大语言模型使用。参与者获得酬金的依据是最终任务产出，而使用大语言模型本身反而需要支付每项任务 0.25 元的成本。因此，实验设计并未直接诱导参与者“为了使用大语言模型而使用大语言模型”。更准确地说，绩效关联更强的薪酬结构提高了改善最终产出的边际收益，而大语言模型采纳是参与者面对更高边际收益时作出的一种行为反应。

（二） 大语言模型接入、薪酬结构与劳动生产率

接下来，本文考察大语言模型接入是否提高劳动生产率，以及这一效应是否随薪酬机制不同而变化。

1. 分酬金规则的估计结果

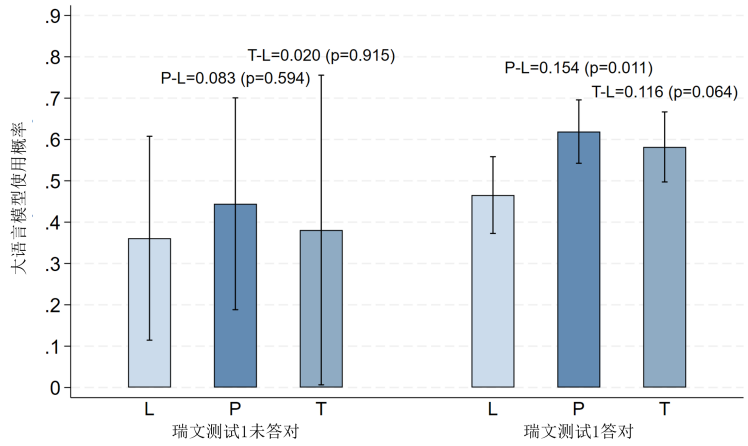
首先，本文在三种酬金规则下分别估计大语言模型接入对劳动生产率的影响。具体估计如下模型：

$$Y_{its} = \rho LLMAccess_i + \Gamma' X_{it} + \lambda_t + \delta_s + \varepsilon_{its}, \tag{2}$$

其中，Y_{its} 表示三类标准化生产率指标之一，包括标准化绩效得分、标准化质量调整后绩效得分和标准化产出数量；$LLMAccess_i$ 表示参与者是否被分配到大语言模型接入组；其他变量含义与式(1)相同。该模型分别在固定支付组、绩效支付组和锦标赛组样本中估计。

图5展示了不同酬金规则下大语言模型接入组和无大语言模型接入组的平均生产率差异。三个面板分别对应标准化绩效得分、标准化质量调整后绩效得分和标准化产出数量。结果显示，在所有酬金规则下，大语言模型接入组的表现均优于无大语言模型接入组，说明大语言模型接入能够显著提高劳动生产率。同时，不同酬金规则下的生产率差距并不相同：固定支付下的差距相对较小，而绩效支付和锦标赛激励下的差距更大。

Panel A. 按瑞文测试 *1* 表现分组（较易题）



Panel B. 按瑞文测试 *2* 表现分组（较难题）

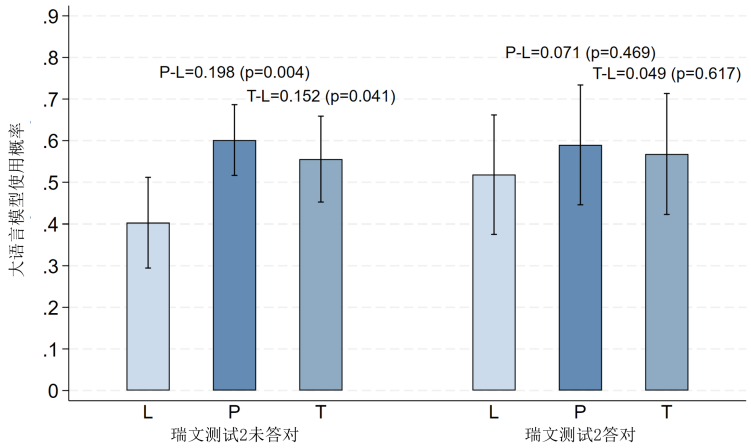


图 4 按认知能力划分的大语言模型采纳异质性

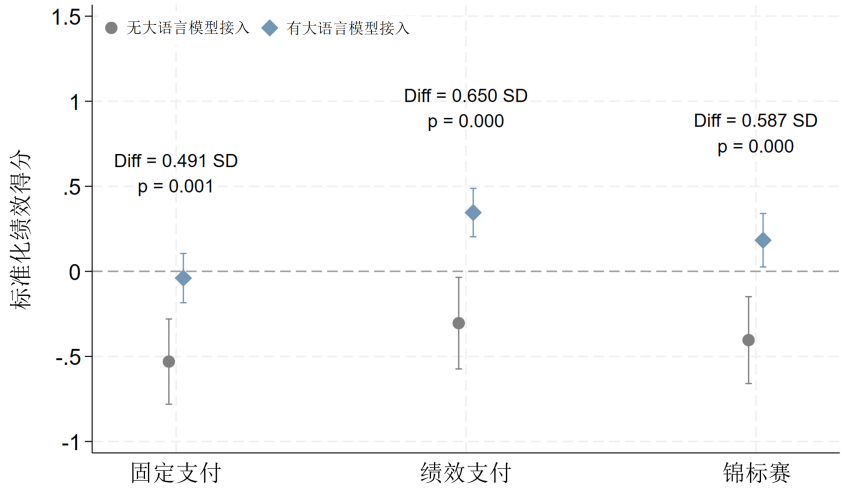
注：本图报告大语言模型接入组在不同薪酬结构下的平均使用概率。Panel A 按参与者是否答对较易的瑞文测试 1 分组，Panel B 按是否答对较难的瑞文测试 2 分组。L、P 和 T 分别表示固定支付、绩效支付和锦标赛。误差线为 95% 置信区间；差异和 *p* 值均相对于固定支付组计算，标准误聚类到参与者层面。

表4进一步量化了上述结果。总体来看，大语言模型接入对三类生产率指标均具有显著正向影响。在使用标准化绩效得分作为因变量时，大语言模型接入的估计系数在 0.407 至 0.806 个标准差之间；在使用标准化质量调整后绩效得分作为因变量时，估计系数在 0.419 至 0.888 个标准差之间；在使用标准化产出数量作为因变量时，估计系数在 0.418 至 0.999 个标准差之间。这说明大语言模型接入不仅提高了产出数量，也提高了质量相关的任务表现。

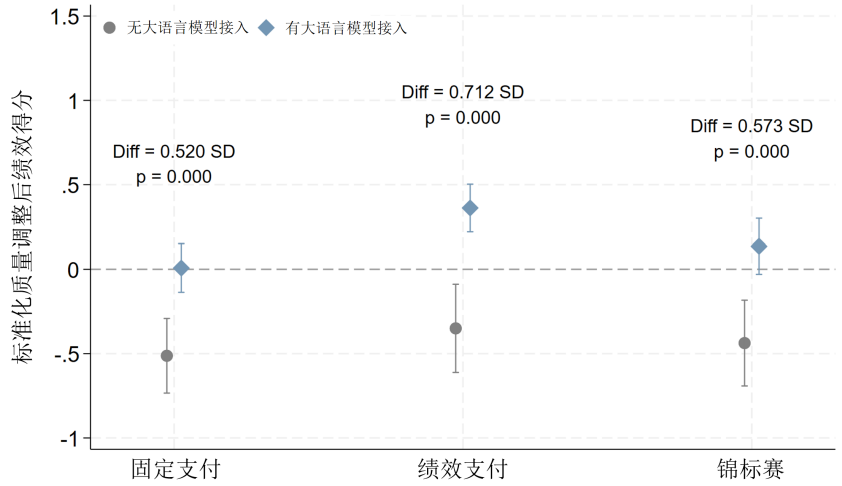
进一步比较不同酬金规则可以发现，大语言模型接入的生产率效应在绩效关联更强的薪酬结构下更大。在固定支付组中，加入控制变量后，大语言模型接入使标准化绩效得分提高 0.407 个标准差，使标准化质量调整后绩效得分提高 0.419 个标准差，使标准化产出数量提高 0.418 个标准差。在绩效支付组中，对应估计值分别为 0.711、0.694 和 0.674 个标准差，均明显高于固定支付组。这表明，当酬金与绝对绩效更紧密相关时，大语言模型接入所带来的生产率收益被进一步放大。

锦标赛组的结果同样表明大语言模型接入具有更大的生产率效应。在加入控制变量后，大语言模型接入使标准化绩效得分提高 0.806 个标准差，使标准化质量调整后绩效得分提高 0.888 个标准差，使标准化产出数量提高 0.999 个标准差。这些估计值不仅高于固定支付组，也高于绩效支付组的对应估计值。由此可见，当酬金取决于相对排名时，大语言模型接入可能带来更大的任务表现提升。

Panel A. 标准化绩效得分



Panel B. 标准化质量调整后绩效得分



Panel C. 标准化产出数量

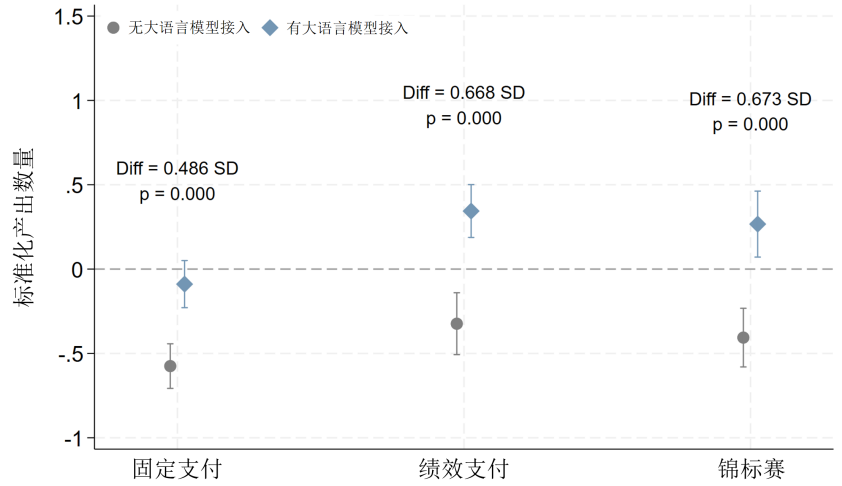


图 5 薪酬结构、大语言模型接入与劳动生产率

注：本图按薪酬结构和大语言模型接入状态报告平均生产率。Panel A 的绩效得分依据以客观表现为核心的评分规则构造；Panel B 的质量调整后绩效得分进一步加入反映答案质量的奖励和惩罚；Panel C 的产出数量在任务 1 和任务 3 中以字数衡量，在任务 2 中以完成的小题数量衡量。所有指标均在任务内标准化。点表示组均值，误差线为 95% 置信区间；图中差异和 p 值为同一薪酬结构内接入组与无接入组之差，标准误聚类到参与者层面。

表 4 薪酬结构、大语言模型接入与劳动生产率

	固定支付		绩效支付		锦标赛		全样本	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. 因变量为标准化绩效得分								
大语言模型接入	0.491***	0.407**	0.650***	0.711***	0.587***	0.806***	0.491***	0.407**
	(0.142)	(0.171)	(0.148)	(0.161)	(0.145)	(0.166)	(0.142)	(0.171)
绩效支付							0.226	-0.500
							(0.177)	(1.199)
锦标赛							0.126	0.921
							(0.172)	(1.169)
大语言模型接入 × 绩效支付							0.159	0.304
							(0.204)	(0.235)
大语言模型接入 × 锦标赛							0.096	0.399*
							(0.202)	(0.237)
R^2	0.066	0.205	0.084	0.232	0.071	0.245	0.091	0.245
Panel B. 因变量为标准化质量调整后绩效得分								
大语言模型接入	0.520***	0.419***	0.712***	0.694***	0.573***	0.888***	0.520***	0.419***
	(0.130)	(0.149)	(0.144)	(0.147)	(0.147)	(0.166)	(0.129)	(0.149)
绩效支付							0.162	-0.967
							(0.165)	(1.183)
锦标赛							0.075	0.155
							(0.162)	(1.214)
大语言模型接入 × 绩效支付							0.192	0.275
							(0.193)	(0.210)
大语言模型接入 × 锦标赛							0.053	0.469**
							(0.195)	(0.222)
R^2	0.069	0.207	0.089	0.223	0.064	0.267	0.093	0.246
Panel C. 因变量为标准化产出数量								
大语言模型接入	0.486***	0.418***	0.668***	0.674***	0.673***	0.999***	0.486***	0.418***
	(0.095)	(0.117)	(0.118)	(0.156)	(0.129)	(0.205)	(0.095)	(0.117)
绩效支付							0.251**	0.231
							(0.109)	(0.998)
锦标赛							0.169	2.632**
							(0.104)	(1.076)
大语言模型接入 × 绩效支付							0.182	0.256
							(0.151)	(0.195)
大语言模型接入 × 锦标赛							0.187	0.581**
							(0.159)	(0.234)
R^2	0.083	0.191	0.102	0.244	0.075	0.293	0.107	0.271
基线控制变量	否	是	否	是	否	是	否	是
实验场次固定效应	否	是	否	是	否	是	否	是
观测值	306	306	294	294	267	267	867	867

注：本表报告大语言模型接入、薪酬结构及其交互项对劳动生产率的 OLS 估计结果。所有模型均控制任务固定效应，标准误聚类到参与者层面。第 (1)–(6) 列在三种酬金规则内分别估计，以无大语言模型接入组为对照；第 (7)–(8) 列的省略组为固定支付且无大语言模型接入组。第 (8) 列允许基线控制变量、任务固定效应和实验场次固定效应的系数随酬金规则变化。三项因变量均在任务内标准化。基线控制变量定义见正文。显著性水平：* $p < 0.1$，** $p < 0.05$，*** $p < 0.01$。

2. 合并样本的交互项估计

为更直接地检验大语言模型接入效应是否随酬金规则变化，本文进一步在全样本中估计交互项模型：

$$\begin{aligned} Y_{its} = {} & \theta_1 LLMAccess_i + \theta_2 PerformanceBased_i + \theta_3 Tournament_i \\ & + \theta_4 (LLMAccess_i \times PerformanceBased_i) + \theta_5 (LLMAccess_i \times Tournament_i) \\ & + \sum_{k \in \{L,P,T\}} \mathbf{1}\{Incentive_i = k\} \left(\Gamma_k' X_{it} + \lambda_{kt} + \delta_{ks}\right) + \varepsilon_{its}, \end{aligned} \tag{3}$$

其中，L、P 和 T 分别表示固定支付、绩效支付和锦标赛激励。该模型中，省略组为固定支付且无大语言模型接入组；在完整设定中，基线控制变量、任务固定效应和实验场次固定效应均允许随酬金规则变化。

表4第 (7) 和第 (8) 列报告了合并样本的交互项估计结果。第 (7) 列模型设定较简约，包含处理变量、交互项和任务固定效应；第 (8) 列为完整设定，加入基线控制变量和实验场次固定效应，并允许控制变量和固定效应的系数在不同酬金规则之间变化。

结果与分组回归基本一致。首先，$LLMAccess_i$ 的系数在三个生产率指标上均为正且显著，表明即使在固定支付这一绩效边际回报较低的薪酬结构下，大语言模型接入仍能提高劳

动生产率。其次，大语言模型接入与绩效支付、锦标赛激励的交互项均为正，说明相较于固定支付组，绩效支付和锦标赛激励下大语言模型接入的生产率效应更大。在较简约设定中，绩效支付交互项通常更大，但未达到常规统计显著水平。这一精度不足可能主要来源于每个实验组内样本量有限。

在加入控制变量和固定效应的完整设定中，锦标赛交互项变得更为显著。对于标准化绩效得分，大语言模型接入与锦标赛激励的交互项为 0.399，并在 10% 水平上显著；对于标准化质量调整后绩效得分，交互项为 0.469，并在 5% 水平上显著；对于标准化产出数量，交互项为 0.581，并在 5% 水平上显著。相比之下，绩效支付交互项仍为正，但统计显著性较弱。该结果表明，在控制参与者特征和实验场次差异后，锦标赛激励下大语言模型接入所带来的生产率增益相对更加突出。

综合来看，大语言模型接入显著提高劳动生产率，但其效应并非在所有薪酬结构下完全相同。当酬金更紧密地取决于绝对绩效或相对绩效时，大语言模型接入的生产率收益更大。这说明大语言模型的实际回报并不是单纯由技术接入决定的，而是与劳动者使用技术时所处的组织薪酬结构密切相关。

（三） 结果解释

上述结果表明，薪酬结构同时影响大语言模型采纳和大语言模型接入的生产率效应。一方面，绩效支付和锦标赛激励提高了参与者使用大语言模型的概率，说明大语言模型使用是对薪酬结构作出的内生行为反应。另一方面，在绩效关联更强的薪酬结构下，大语言模型接入对劳动生产率的影响更大，说明同一项技术在不同组织环境下可能产生不同的实际回报。

这一结果可能通过两个渠道产生。第一，绩效关联更强的薪酬结构提高了大语言模型采纳概率，使更多参与者使用这一有效工具。第二，绩效关联更强的薪酬结构可能不仅影响是否使用大语言模型，也影响使用后的交互质量。例如，参与者在绩效关联更强的薪酬结构下可能提出更具体的问题、进行更多轮追问，或更认真地修改和整合大语言模型输出。这一使用强度和交互方式的渠道十分重要，但是本文的识别重点在于“薪酬结构—大语言模型采纳—劳动生产率”这一可由随机设计和任务层面使用记录直接支持的路径。

本文的核心结果可以理解为：大语言模型的生产率回报具有组织依赖性。同一项大语言模型技术是否能够转化为生产率提升，不仅取决于劳动者是否拥有技术接入机会，也取决于在特定激励结构下劳动者是否选择使用技术并将其整合进自身工作当中。

五、 机制分析

（一） 来自开放式回答的启示

主要结果表明，绩效支付和锦标赛激励均会提高大语言模型使用率，且绩效关联更强的薪酬结构会放大大语言模型接入的生产率效应。不过，两类薪酬结构呈现出的模式并不完全相同：绩效支付对平均采纳率的促进作用通常更大、更稳定；而在加入控制变量后，锦标赛下大语言模型接入的生产率增益更为突出。这一差异提示，绩效支付和锦标赛可能通过不同的行为机制发挥作用。以下分析尝试解释核心结果并提供提示性机制证据。

为理解不同薪酬结构下可能的机制差异，三项任务结束后，部分大语言模型接入组参与者回答了一道开放式问题，说明在其面对的酬金规则下，大语言模型是否有助于获得更高酬金，以及为什么在三项任务中选择使用或不使用大语言模型。这些回答揭示了参与者如何理解薪酬结构并权衡大语言模型使用的成本和收益，为下文的机制框架提供动机。

固定支付组的参与者经常认为使用大语言模型的激励较弱，因为该规则只要求达到最低合格标准，而不要求最大化任务表现。一名参与者写道：

> “我认为 AI 助手不能帮助我获得更多酬金，因为每个题目我认为我的答案能够达到最低标准，不是特别困难。在前面的任务中，我不用 AI 助手一方面是因为会减少奖励，一方面是我觉得我有能力完成任务。”

即使选择使用大语言模型，一些参与者也并非主要将其视为提高绩效的工具，而是把它用于降低努力成本、快速生成合格答案。另一名使用了大语言模型助手的参与者解释道：

> “由于抽筋[6]发放采用的是及格即可获得定额酬金的二分类酬金获得法，因此我不需要不断优化回答质量，而仅需快速给出一个合格可用的答案即可。”

[6]参与者提交的原文如此，应为“酬金”。

这些回答表明，在固定支付下，参与者通常不会从边际绩效提升的角度评价大语言模型；更相关的考虑是，在已经能够达到最低标准的情况下，大语言模型能否降低完成任务的努力成本。

相比之下，绩效支付组的参与者更常强调大语言模型改善产出质量和节省时间的作用。一名参与者写道：

> “附加酬金是按答题质量和时间综合评定，质量越高酬金越多，每题 5r，每次用 ai 会扣除 0.25r，但能节省时间、大大提高答题效率。……第三个任务分配资金，ai 给出的方案逻辑清晰合理，更倾向于客观，所以使用 ai 生成分配方案并稍加修改即可。ai 与人工添改的结合能大大提高效率节省时间，相比纯人工我觉得也提升了质量，但意见出现分歧时我会更相信自己的判断。”

这一回答与如下解释一致：在绩效支付下，参与者把大语言模型视为提高绝对绩效的工具；只要它能改善最终产出或减少时间成本，就能提高预期报酬。

锦标赛组的回答体现了不同逻辑。一些参与者不仅关心大语言模型能否提高绝对表现，还关心它能否帮助自己超越他人。一名参与者写道：

> “AI 可能是观察大部分人的行为选择取频率较高的作为‘保险’回答，可能更全面，但我认为和我的想法会产生较大偏差。10 个人的小样本下方差应该挺大的。我应该自己作答显现差异。”

该回答说明，在锦标赛下，大语言模型不仅是一项生产工具，也是一项策略性投入：它的价值取决于能否形成相对优势，还是会使不同参与者的答案更加同质。

总体而言，开放式回答首先表明参与者的行为与其面对的薪酬结构大体一致。固定支付组会理性权衡使用成本与最低合格标准；绩效支付组更关注大语言模型能否提高绝对表现；锦标赛组则额外关注它是否有助于差异化竞争。基于这些回答，本文将后两类薪酬结构的作用渠道概括为“绩效提升机制”和“差异化机制”。

（二） 绩效提升机制与差异化机制

第一类机制是绩效提升机制。根据这一机制，劳动者使用大语言模型，是因为绩效关联更强的薪酬结构提高了任何能够改善客观任务表现的工具的价值。如果劳动者认为大语言模型有助于提高答案质量、完成速度或产出数量，那么绩效关联更强的薪酬结构应当提高大语言模型使用率。在这一机制下，大语言模型的价值主要来自绝对绩效改善，因此该机制与绩效支付最为相关。

第二类机制是差异化机制。根据这一机制，劳动者使用大语言模型，是因为他们预期大语言模型能够帮助自己相对于竞争者形成优势。如果大语言模型产生的答案能够帮助劳动者更好地组织思路、覆盖更多要点或提出不同于他人的方案，那么大语言模型可能提升其相对排名。然而，如果劳动者认为大语言模型会使不同使用者的答案更加相似，那么依赖大语言模型反而可能降低其差异化程度。在这种情况下，锦标赛激励未必总是提高大语言模型使用率；在强调原创性、独特性或相对表现的任务中，这一机制预测锦标赛激励甚至可能抑制大语言模型使用。

这一区分意味着，绩效支付和锦标赛激励对大语言模型采纳的影响可能具有不同模式。绩效支付应当主要在劳动者认为大语言模型能够提高绝对表现时提高大语言模型使用率；锦标赛激励的影响则更具选择性，只有当劳动者认为大语言模型能够带来相对竞争优势时，锦标赛激励才更可能促进大语言模型采纳。

（三） 来自任务差异的证据

首先，本文利用不同任务之间的差异考察上述机制。图3的分任务结果显示，薪酬结构对大语言模型采纳的影响并不在三项任务中完全相同。为更系统地刻画这一差异，本文在每项任务内分别估计如下模型：

$$LLMUse_{i\tau s} = \beta_{1\tau} PerformanceBased_i + \beta_{2\tau} Tournament_i + \Gamma_\tau' X_{i\tau} + \delta_s + \varepsilon_{i\tau s}, \quad (4)$$

其中，$\tau \in \{1,2,3\}$ 表示任务编号，其他变量定义与前文相同。结果对应于表3的第 (3)–(8) 列。

任务 2 中大语言模型采纳对薪酬结构的反应最弱。该任务要求参与者在每个视频案例中从 A、B 两个选项中选择更可能获得更高播放量的视频，并简要说明理由。由于任务形式相对简单，且大语言模型未必具备判断具体视频热度所需的背景知识，参与者可能认为大语言模型难以带来明显帮助。因此，即使酬金规则提供了更高的绩效边际回报，参与者也没有显著提

高大语言模型使用率。这一结果说明，绩效关联更强的薪酬结构并不会在所有任务中机械地提高大语言模型使用。

任务 1 和任务 3 的结果更能反映两类机制之间的差异。任务 1 是新闻稿修改与编辑，与大语言模型的语言处理能力较为契合。在该任务中，绩效支付和锦标赛激励均显著提高大语言模型使用率。这说明当大语言模型被认为能够提高任务质量和效率时，两类绩效关联更强的薪酬结构都可能促进大语言模型采纳。任务 3 是预算分配任务，既需要结构化表达，也需要判断、取舍和方案说明。在该任务中，绩效支付显著提高大语言模型使用率，而锦标赛激励的影响较弱。这一模式与差异化机制相一致：在需要判断和解释的任务中，参与者可能担心大语言模型生成的方案过于常规或与其他竞争者相似，从而降低其在锦标赛中的相对优势。

因此，分任务结果支持本文的机制解释。绩效支付主要通过提高绝对绩效回报，使参与者更愿意使用可能改善产出的大语言模型工具；锦标赛激励则取决于任务是否允许大语言模型帮助参与者形成相对优势。当大语言模型对任务适配度较低，或其输出可能降低答案差异化时，锦标赛激励对大语言模型采纳的促进作用会减弱。

（四） 来自大语言模型绩效预期的证据

其次，本文利用参与者对自身表现和大语言模型表现的评价进一步考察绩效提升机制。每项任务结束后，参与者分别评价自己和大语言模型在该任务中的预期表现。本文据此将样本分为两组：一组为参与者认为大语言模型表现低于自身表现，另一组为参与者认为大语言模型表现不低于自身表现。

图3的 Panel C 表明，当参与者认为大语言模型的表现低于自己时，绩效支付和锦标赛均未显著提高大语言模型使用率；在锦标赛组中，点估计甚至与大语言模型采纳率下降相一致。相比之下，当参与者认为大语言模型的表现将优于自己，或至少不差于自己时，绩效支付和锦标赛均显著提高了大语言模型使用率。这一结果表明，薪酬结构主要在劳动者认为大语言模型确实具有绩效优势时才会显著影响其采纳决策。

这一异质性结果与前文提出的两种机制相吻合，但其可能的作用方式有所不同。对于绩效支付而言，如果劳动者预期大语言模型能够提高任务产出，那么更高的绩效回报会提高使用大语言模型的收益，从而使其更具吸引力。对于锦标赛而言，当大语言模型被认为表现较弱时，锦标赛并不会为采纳大语言模型提供明显的额外动机，甚至可能强化参与者依靠自身判断或原创性进行竞争的倾向，从而抑制大语言模型的采纳。相反，当大语言模型被认为表现较强时，锦标赛能够提高其使用率，因为此时大语言模型可能成为获得相对竞争优势的来源。

（五） 来自同伴大语言模型使用信念的证据

第三，本文利用参与者对同伴大语言模型使用率的信念进一步检验差异化机制。任务后问卷要求大语言模型接入组参与者预测其他参与者使用大语言模型的比例。基于该变量，本文考察当参与者认为其他人更可能或更不可能使用大语言模型时，不同薪酬机制对自身大语言模型采纳的影响是否发生变化。

图3的 Panel D 表明，当参与者认为其他人较有可能使用大语言模型时，绩效支付仍然显著提高大语言模型使用率。这一结果与绩效提升机制相一致。如果参与者将大语言模型视为一种能够改善绝对任务表现的有效工具，那么其他人是否也使用大语言模型，并不会降低其自身使用该工具的价值。甚至从社会学习的角度看，其他人普遍使用大语言模型还可能进一步强化个体采纳该工具的动机。

锦标赛激励则呈现出不同的结果。当参与者认为其他人较有可能使用大语言模型时，锦标赛对大语言模型使用的影响明显减弱，并且往往不再显著。这一结果正是差异化机制所预期的。如果竞争对手普遍使用大语言模型，那么使用大语言模型便不太可能带来相对优势；相反，它可能只是使不同参与者的回答更加趋同，而在以超越他人为奖励依据的锦标赛环境中，这种趋同并不具有吸引力。

当参与者认为其他人不太可能使用大语言模型时，上述逻辑则发生反转。在这种情况下，绩效支付对大语言模型使用的促进作用减弱。这与这样一种情形相一致：大语言模型并未被参与者普遍视为一种能够可靠提高绝对绩效的工具。相比之下，锦标赛对大语言模型使用的影响变得更加正向，因为此时大语言模型可能成为参与者使自身回答区别于竞争对手的一种手段。换言之，当大语言模型并非大多数参与者都会采用的常规策略时，锦标赛反而可能提高其吸引力，因为它为参与者提供了一种采取非对称竞争策略的可能方式。

这一分组分析较为清晰地区分了两种机制。绩效支付主要在大语言模型被认为有助于提高绝对任务表现时促进其采纳，而这种作用并不取决于其他人是否也使用大语言模型。相比之下，锦标赛主要在大语言模型能够帮助参与者形成相对竞争优势时促进其采纳；当参与者

预期大语言模型会使竞争者之间的回答更加趋同时，这一作用便会减弱。

（六） 机制解释

因此，图3和表3中的证据表明，绩效支付和锦标赛可能通过不同渠道提高大语言模型使用率。绩效支付主要通过提高那些被认为能够改善绝对绩效的工具的使用价值发挥作用。相比之下，锦标赛的作用更具策略性：只有当大语言模型被预期能够帮助参与者取得相对于他人的优势时，锦标赛才会促进其使用；而当大语言模型可能导致不同参与者的产出趋于同质化时，这一作用便会减弱。

这一区分也有助于理解前文的主要结果。在大语言模型采纳回归中，绩效支付对大语言模型使用的促进作用通常比锦标赛更大，也更加稳定。然而，在表4合并样本的生产率回归中，加入控制变量后，锦标赛与大语言模型接入的交互项反而更加突出。这两组结果并不矛盾。绩效支付可能在更广泛的参与者群体中提高大语言模型使用率，而锦标赛则可能更加有选择地促使参与者在大语言模型能够带来最大相对绩效优势的情形下使用该工具。因此，即使锦标赛并未在平均采纳率上产生最大的影响，它仍可能显著放大大语言模型接入所带来的生产率增益。

更一般地说，机制分析表明，薪酬结构与大语言模型采纳之间的关系不能简单概括为“更强的激励会导致更多的大语言模型使用”。绩效关联更强的薪酬结构是否会提高大语言模型采纳，取决于劳动者面对何种薪酬机制、其如何判断大语言模型在当前任务中的有效性，以及大语言模型究竟有助于改善绝对绩效，还是有助于形成相对差异化优势。区分这两种渠道，对于理解为什么同一种大语言模型技术在不同薪酬制度下可能产生不同的生产率后果至关重要。

上述机制解释对组织激励设计具有启示意义。当大语言模型主要改善客观质量、效率或产出数量时，绩效支付可能更有效地促进生产性采纳。当工作任务更强调差异化、原创性或相对表现时，锦标赛激励可能带来更高回报，但其前提是大语言模型能够帮助劳动者形成差异化优势，而不是使劳动者产出趋于同质化。

六、 结论

本文研究薪酬结构如何影响劳动者对大语言模型的采纳，以及大语言模型接入在不同薪酬结构下如何影响劳动生产率。通过同时随机改变大语言模型接入和酬金规则的实验室实验，本文表明，大语言模型并不是一项在可得后便自动发挥作用的机械性技术冲击。劳动者是否选择使用它，以及大语言模型接入能够带来多大的生产率收益，都取决于其面对的薪酬结构。

本文得到三个主要结论。第一，相较固定支付，绩效支付和锦标赛均提高参与者在能够接入大语言模型时选择使用它的概率。第二，大语言模型接入显著提高劳动生产率，这一结果同时体现在绩效得分、质量调整后绩效得分和产出数量上。第三，大语言模型接入带来的生产率提升在绩效支付和锦标赛下大于固定支付，说明技术接入的产出效应并不独立于薪酬合同环境。

机制证据进一步表明，绩效支付和锦标赛并非以相同方式影响大语言模型使用。绩效支付主要通过提高绝对绩效改进的回报，鼓励参与者使用其认为能够提高任务表现的工具。锦标赛的作用则更具选择性，并与相对差异化密切相关：当参与者预期大语言模型能够形成竞争优势时，锦标赛会促进采纳；当参与者担心模型输出使答案趋同时，这一作用会减弱。这一区分有助于解释两类薪酬结构为何会在平均采纳率和大语言模型接入的生产率收益上呈现不同模式。

上述结果对组织管理和学术研究均有启示。对企业而言，部署大语言模型不能仅停留在提供工具接入。同一项技术在不同薪酬结构下可能产生显著不同的实际结果，组织需要将技术部署和薪酬设计共同考虑。对研究者而言，本文强调应将大语言模型使用视为内生选择，而非外生接入后的必然结果。大语言模型的劳动市场效应不仅取决于技术发展水平，也取决于劳动者在何种制度条件下决定是否使用以及如何整合该技术。

本文结论存在若干适用边界。第一，实验在受控实验室中进行，参与者主要为大学生。线下实验室环境有利于保证处理实施和采纳记录的内部有效性，但无法完全覆盖真实企业中劳动者、任务和组织环境的多样性。本文结论最直接适用于短时、认知密集且能够依据明确规则评价产出的任务，不宜机械推广到体力任务、团队生产、难以评价质量的工作，或高度开放且主要依赖原创判断的任务。

第二，本文考察的平台内置大语言模型仅具有较小的单任务进入成本。现实工作中的采纳还可能受到组织规范、合规要求、数据安全、学习成本以及同事和管理者预期等因素影响，这些因素可能强化、削弱甚至改变本文观察到的模式。

第三，本文识别的是短期任务表现，而非长期就业和一般均衡后果。实验没有回答反复使用大语言模型是否会改变劳动者技能、任务分工或组织实践，也不能据此推断技术在更长时期内对岗位和工资的影响。

第四，本文的锦标赛是一种适度、匿名的相对绩效安排。现实中更强或更显著的排名竞争可能诱发非线性甚至扭曲性反应，例如过度依赖大语言模型、策略性差异化、答案同质化，或忽视评分规则未充分覆盖的质量维度。因此，本文结果不意味着竞争强度越高，大语言模型辅助下的生产率就一定越高。

最后，本文主要识别任务层面的采纳扩展边际，并利用任务差异和实验后信念提供提示性机制证据。更细致的使用方式、长期学习和组织调整超出本文范围。将这一框架拓展到真实工作场景、团队生产和长期组织环境，是颇有价值的未来研究方向。

总而言之，本文强调了一个直观但重要的思想：大语言模型的生产率后果不仅由技术本身决定，也由组织工作的经济制度决定。理解大语言模型如何影响劳动市场，必须同时理解劳动者为什么选择使用或不使用它。换言之，大语言模型在工作场景中的未来前景，不仅取决于“技术能够做什么”，更取决于“组织如何为技术所带来的生产率提升进行付酬”。

参考文献

[1] Acemoglu, D., and P. Restrepo, "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment", *American Economic Review*, 2018, 108(6), 1488–1542.

[2] Acemoglu, D., and P. Restrepo, "Robots and Jobs: Evidence from US Labor Markets", *Journal of Political Economy*, 2020, 128(6), 2188–2244.

[3] Acemoglu, D., "The Simple Macroeconomics of AI", *Economic Policy*, 2025, 40(121), 13–58.

[4] Autor, D. H., F. Levy, and R. J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration", *The Quarterly Journal of Economics*, 2003, 118(4), 1279–1333.

[5] Baek, C., T. Tate, and M. Warschauer, "ChatGPT Seems Too Good to Be True: College Students' Use and Perceptions of Generative AI", *Computers and Education: Artificial Intelligence*, 2024, 7, 100294.

[6] Bandiera, O., I. Barankay, and I. Rasul, "Social Preferences and the Response to Incentives: Evidence from Personnel Data", *The Quarterly Journal of Economics*, 2005, 120(3), 917–962.

[7] Bandiera, O., I. Barankay, and I. Rasul, "Incentives for Managers and Inequality among Workers: Evidence from a Firm-Level Experiment", *The Quarterly Journal of Economics*, 2007, 122(2), 729–773.

[8] Bandiera, O., I. Barankay, and I. Rasul, "Social Incentives in the Workplace", *The Review of Economic Studies*, 2010, 77(2), 417–458.

[9] Bandiera, O., I. Barankay, and I. Rasul, "Team Incentives: Evidence from a Firm Level Experiment", *Journal of the European Economic Association*, 2013, 11(5), 1079–1114.

[10] Bick, A., A. Blandin, and D. J. Deming, "The Rapid Adoption of Generative AI", NBER Working Paper No. 32966, 2025.

[11] Bresnahan, T. F., E. Brynjolfsson, and L. M. Hitt, "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence", *The Quarterly Journal of Economics*, 2002, 117(1), 339–376.

[12] Brynjolfsson, E., D. Li, and L. Raymond, "Generative AI at Work", *The Quarterly Journal of Economics*, 2025, 140(2), 889–942.

[13] Chatterji, A., T. Cunningham, D. J. Deming, Z. Hitzig, C. Ong, C. Y. Shan, and K. Wadman, "How People Use ChatGPT", NBER Working Paper No. 34255, 2025.

[14] Cui, K. Z., M. Demirer, S. Jaffe, L. Musolff, S. Peng, and T. Salz, "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers", *Management Science*, 2026.

[15] Dell'Acqua, F., E. McFowland III, E. Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality", Harvard Business School Working Paper No. 24-013, 2023.

[16] Gibbons, R., "Incentives in Organizations", *Journal of Economic Perspectives*, 1998, 12(4), 115–132.

[17] Greiner, B., P. Grünwald, T. Lindner, G. Lintner, and M. Wiernsperger, "Incentives, Framing, and Reliance on Algorithmic Advice: An Experimental Study", *Management Science*, 2026, 72(1), 302–322.

[18] Holmstrom, B., and P. Milgrom, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design", *Journal of Law, Economics, and Organization*, 1991, 7(Special Issue), 24–52.

[19] Humlum, A., and E. Vestergaard, "The Unequal Adoption of ChatGPT Exacerbates Existing Inequalities among Workers", *Proceedings of the National Academy of Sciences*, 2025, 122(1), e2414972121.

[20] Lazear, E. P., "Performance Pay and Productivity", *American Economic Review*, 2000, 90(5), 1346–1361.

[21] Lazear, E. P., and S. Rosen, "Rank-Order Tournaments as Optimum Labor Contracts", *Journal*

of Political Economy, 1981, 89(5), 841–864.
[22] Lazear, E. P., and K. L. Shaw, "Personnel Economics: The Economist's View of Human Resources", *Journal of Economic Perspectives*, 2007, 21(4), 91–114.
[23] Noy, S., and W. Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence", *Science*, 2023, 381(6654), 187–192.
[24] Prendergast, C., "The Provision of Incentives in Firms", *Journal of Economic Literature*, 1999, 37(1), 7–63.
[25] Shearer, B., "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment", *The Review of Economic Studies*, 2004, 71(2), 513–534.

附录 A 实验界面

Panel A：大语言模型接入组界面

Panel B：无大语言模型接入组界面

图 A1 不同大语言模型接入状态下的实验界面

注：Panel A 展示大语言模型接入组的实验网页界面示例，Panel B 展示无接入组的界面。除大语言模型交互模块外，两类界面保持一致。