arXiv:2304.10014v1 [physics.ed-ph] 19 Apr 2023

利用 ChatGPT 开发未来物理教师的物理任务

Stefan K¨uchemann[1],* Steffen Steinert[1], Natalia Revenga[1], Matthias Schweinberger[1], Yavuz Dinc[1], Karina E. Avila[2], 和 Jochen Kuhn[11] *Chair of Physics Education Research, Faculty of Physics, Ludwig-Maximilians-University Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany and*[2] *Department of Mathematics, RPTU Kaiserslautern-Landau, Paul-Ehrlich-Str. 14, 67663 Kaiserslautern, Germany*（日期：2023年4月21日）

大型语言模型的近期发展为教学与学习带来了诸多机遇。尽管公众对大型语言模型的使用展开了广泛讨论，但关于其在教育领域机遇与风险的实证研究仍然有限。在本研究中，我们展示了准教师使用 ChatGPT 3.5 开发物理任务的优势与不足。在一项随机对照试验中，26名未来物理教师学生被分为两组：第一组使用 ChatGPT 3.5，为高中十年级学生开发运动学领域四个不同概念的基于文本的物理任务；第二组则使用传统教科书，为相同概念和目标组编制任务。结果显示，两组在任务正确性方面没有差异，但使用教科书的学生所编任务清晰度更高，也更常将问题置于有意义的情境中。两组都能轻松地将任务难度调整到符合目标组的水平，但在确保任务具有足够的具体性方面都遇到了很大困难，也就是说，解决任务所需的相关信息有所缺失。使用 ChatGPT 提出问题的学生对系统可用性的评价较高，但在输出质量方面遇到了困难。这些结果揭示了在教育中使用大型语言模型的机遇与隐患。

一、引言

任务开发是学校教师的一项必备技能，尤其是在设计形成性与总结性评价时。高质量的任务有助于教师诊断学生的概念理解及其困难、监测学习进展，并评估教学方法的有效性。借助精心设计的评估，教师能够及时采取干预措施，支持学生发展，进而直接影响学生表现。通过设计和实施多样化的任务，以适应不同的学习偏好和先前知识水平，教师能够营造包容性课堂，兼顾学生的多样性，从而直接影响教育质量。因此，1990 [1], 制定了学生评价方面的教师能力标准，并将该能力认定为学科教学知识 [2]和课堂评价 [3]的重要组成部分。尽管其相关性已得到认可，研究表明，美国教师认为自己在开发用于评估学生表现的任务方面准备不足 [4, 5], ，即使他们接受过本科阶段的评价培训也是如此。这种看法与有关评价素养有限的研究发现相符 [6]。因此，亟需确定培训未来教师和在职教师开发任务的方法，尤其是借助数字工具开展培训 [7]。

除了许多其他应用之外，ChatGPT 等大型语言模型的最新进展，也为利用精心设计的提示词自动生成评估内容提供了可能性 [8]。这引发了关于准教师使用 ChatGPT 开发的任务，与借助教科书等传统方法开发的任务相比，质量如何的问题。通过让教师指定提示词和任务创建标准，ChatGPT 或可简化并改进任务开发过程。

本研究比较了未来物理教师使用 ChatGPT（干预组）与使用教科书（对照组）设计的任务质量。研究聚焦于以下三个主要问题：

RQ1：未来物理教师使用 ChatGPT 设计的物理任务，与使用教科书设计的任务相比，质量如何?

RQ2：未来物理教师对 ChatGPT 生成的任务和教科书中的任务做出了哪些改进?

RQ3：物理教师在设计物理任务时，如何评价 ChatGPT 的可用性、感知有用性和输出?

下文将通过概述高中物理评估、教师设计作业的能力，以及教育领域中大型语言模型迅速发展的研究现状，为我们的研究问题奠定基础。我们将聚焦于第二节 A 部分所概述的参数，考察使用 ChatGPT 是否能提升准教师所设计的物理评估任务的质量。

*mailto: s.kuechemann@lmu.de

II. 理论背景

A. 高中物理学评估

物理学教育从根本上说，涉及运用和解答各类任务，这些任务有助于组织学生的学习，使教师能够监测学生的学习进展，并为衡量学习表现提供手段 [9]。尽管书面表现评估常被视为一种繁重的前置要求，解答任务对于理解物理学仍至关重要 [10]。学生必须完成大量任务才能成功理解物理学，这要求他们区分学习任务与表现任务 [9], ，并了解何时可以容许错误 [11, 12]。任务可以口头或书面形式呈现，而测验任务主要采用书面形式。基于文本的作业要求学生具备阅读能力，能够运用物理概念和数学方法，并进行批判性反思 [13]。

设计有效任务或根据课堂条件的变化对其加以调整，是一项富有创造性且需要反复迭代的工作。它包括不断尝试、总结经验并修改任务的循环过程。 [9]。考试任务应评估个人能力；选择题虽属合适形式，但可能仅涉及单一维度。这类单维度任务是常规考试文化的典型特征，并可通过统计分析清晰呈现学生的能力概况 [9]。然而，测试中使用的文本必须表述明确，以免产生误解。此类任务中的失误通常归因于对文本的误读，或知识不足、基础理解不充分 [9]。传统任务往往忽视学生的学习过程，从而阻碍稳固概念的形成 [14]。这类任务主要评估阅读或数学能力，而非物理概念思维能力。

在美国 [15] 和德国 [16, 17] 等国家实施基于能力的学习标准和课程 [10]，需要进一步发展任务文化。以学生能力和发展为重点的任务，可以通过将物理学内容的元结构（基本概念和指导思想）以及需要学习的物理学能力应用于实例并鼓励反思，使其更加明晰 [9]。基于能力的任务通常具有鲜明的情境导向，且往往会先提供包含相关和无关信息的文本或材料，以构建情境框架。图片、图表或报纸剪报等视觉材料也可作为导入的一部分。之后，任务会通过命令动词 [18] 提出，并兼顾评价和交流等子能力。这些任务可能包含处理和展示的方法、处理时的辅助材料和提示、有关展示的信息，以及处理和评价标准。最具挑战性的任务是开放式实验综合问题解决任务 [19], ，它们允许存在多个合理解法，也可能没有唯一明确的答案。对于开放式任务，通常必须先明确其所依据的问题，才能解决问题。

任务是物理学课程及相关学习评估的重要组成部分。尽管任务经常使用，但选择和设计合适的任务不容低估。任务必须有效体现预期的学习目标，吸引通常由不同类型学生组成的学习群体，并兼顾学生对物理概念的个体理解 [9]。

B. 教师创建评估的能力

关于教师为有效促进学生学习所需具备的教学法技能和学科专业技能，已有大量研究，并据此提出了若干描述教师基本能力的模型 [20, 21]。近期数学和科学领域的研究表明，教师的学科教学知识（PCK）会显著影响学生的学业成就 [22]。Shulman 强调了学科教学知识的重要性，它使教师能够将学科知识与教学技能相结合，从而让学生更容易理解学科内容 [21, 23, 24]。学科知识以教育工作者具备问题解决技能为前提。然而，要通过选择、调整或开发任务，将这些技能有效运用于教学目的，教育工作者必须透彻理解任务的教学法分析 [25]。

教育工作者应具备设计和重新表述问题的能力，以便为学生营造有意义的学习环境 [26]。Silver 将这一活动称为问题提出 [27]，它既包括生成新问题，也包括重新表述现有任务。问题提出对学生和教育工作者都至关重要，因为它能够培养学生的创造性思维 [28] ，并帮助教育工作者了解学生的思维过程和概念理解 [29]。例如，教育工作者通过使用问题提出测试和学生自行提出的问题，成功识别出了学生在数学方面的概念误解 [30]。

教育工作者应用和教授问题提出的能力，会对学生的问题解决概念理解产生显著影响。因此，教育工作者必须具备必要的技能和知识，以便设计和重新表述问题，从而为学生开展类似的学习活动 [26]。Lowrie（2002年）强调了问题提出与问题解决作为认知过程之间的密切相关性 [31]。因此，问题提出构成了教育工作者职责中的关键组成部分，因为他们不仅需要向学生呈现问题，还需要引导学生逐步成长为能够独立熟练提出问题的人 [32]。

尽管问题提出十分重要，但它也面临诸多挑

战。例如，Crespo 和 Sinclair（2008年）发现，准教师由于在学生和教育工作者的身份中都不熟悉问题提出，因此难以参与其中 [33]。此外，正如 Cai 和 Hwang（2002年） [34] 以及 Silver 和 Cai（1996年） [35]所表明的那样，所提出问题的质量并非总是很高。这些发现凸显了进一步研究的必要性：教育工作者如何学习将问题提出融入教学，以及如何提高自身问题集的质量，从而有效地将其融入教学 [36]。

C. 教育中的大语言模型

本研究中使用的 ChatGPT 系统 [37] 基于大语言模型（LLM）。大语言模型是用于自然语言处理（NLP）的神经网络，经过大量文本数据集训练，能够针对各种语言相关任务生成类似人类的文本 [8]。GPT [38], 、BERT [39], 和 RoBERTa [40] 等模型彻底改变了自然语言处理领域，并拓展了研究与应用的可能性。在教育领域，这些模型可与聊天机器人集成，为学生创造适应性和个性化的学习体验，同时支持教育工作者履行其职责 [41, 42]。大语言模型已被用于多种方式帮助学生，包括内容生成 [43, 44], 、提升提问技能 [45], 、生成代码解释 [46], 、自动化评估 [47], 以及为语言学习提供反馈 [48]。聊天机器人可以充当对话伙伴，支持那些 испыты外语焦虑 [49] 或沟通准备度较低的学生 [50]。集成大语言模型与聊天机器人有助于打造更具吸引力的学习体验，并帮助学生表达好奇心、理解复杂概念 [45]。研究表明，教育工作者对教育中的人工智能持积极态度 [51], ；感知有用性、易用性以及对基于人工智能的工具的信任等因素会影响他们对这些工具的接受度 [52]。

尽管人工智能和聊天机器人在教育领域具有潜力，但仍有一些挑战和未解决的问题需要应对。要负责任地将人工智能融入教育，需要教育工作者、研究人员和政策制定者等不同群体共同努力 [53]。还需要开展进一步研究，以考察大语言模型和聊天机器人在不同教育情境中的有效性，评估其对学习成果的影响，并探讨潜在的伦理问题和偏见 [8]。

本研究评估的 ChatGPT 版本于 1 月 30^{th}发布，存在一些局限。该版本基于 GPT-3.5，无法处理图像。根据该版本自身提供的信息，在本研究开展之前，它存在若干局限，包括缺乏常识知识、可能生成带有偏见的输出、对话语境有限，以及抽象推理和创造力方面存在困难 [54]。

在2月27日进一步询问有关基于GPT-3.5的ChatGPT版本的局限性时，还报告了其他局限性。这些局限性包括知识存在时间限制、回答可能含糊不清或不明确、生成看似合理但不正确或具有误导性的答案、对输入措辞敏感、回答冗长、可能产生不当内容、缺乏常识、无法提出澄清性问题，以及对较长对话语境的考量有限 [55]。

研究结束后，一个基于 GPT-4 的 ChatGPT 版本开放测试 [56, 57]„ 该版本提供了相关信息，说明其局限与基于 GPT-3.5 的版本有哪些不同。基于 GPT-4 和 GPT-3.5 的 ChatGPT 版本在局限方面大体相似，因为这两种模型在设计和训练上存在共同问题。不过，与 GPT-3.5 相比，GPT-4 有望提供更新的信息、给出更高质量的回答、更有效地处理含糊的问题、更好地保持情境连贯性，并可能更少生成不当内容 [55]。

据宣布，GPT-4 将能够处理图像 [56], ，为教育应用开辟新的可能性。

大语言模型和聊天机器人有望通过提供适应性和个性化的学习体验，彻底改变教育。这些技术可以帮助学习者获取知识，并为教师履行职责提供支持，从而提升教育的参与度和成效。然而，我们必须进一步探究这些技术的能力，并应对其整合所带来的挑战，以确保它们在教育环境中的使用负责任且符合伦理。

III. 方法

A. 参与者

本研究旨在考察聊天机器人在协助物理教育者为高中生开发合适的评估任务方面的潜力。鉴于任务设计相关的挑战（见第 IIB 节）通常与经验有关，本研究重点关注不定期教授物理学的本科预备物理教师学生和研究生预备物理教师。

本研究共有26名来自慕尼黑路德维希-马克西米利安大学的未来物理教师参与，其中女性13名、男性13名，年龄中位数为23岁。在这些参与者中，80%的教学经验不超过一年；该群体中有95%（占总样本的76%）报告称自己几乎没有或完全没有编制评估任务的经验。仅有15%的参与者报告称自己在这方面有一定经验，而只有7%称自己经验丰富，其教学经验为3.5至6年。参与完全自愿，且没有报酬。在

研究开始前，当地伦理委员会审查并批准了研究的开展。

B. 设计与材料

在本研究中，参与者的任务是为德国高中十年级学生（15—16岁）开发四个概念性任务，以评估牛顿力学中的常见概念。这些概念包括速度与加速度之间的关系，以及第一、第二和第三运动定律。这些概念之所以被选中，是因为它们在高中物理课程中具有基础性的重要地位。一半的参与者使用 ChatGTP 3.5 [37]（基于 ChatGTP 3.5 的 1 月 30^{th}版本）作为支持来创建物理任务（干预组），而另一半参与者则可以使用标准高中物理电子教材作为支持来创建物理任务（对照组）[58]。高中物理教科书通常包含教师可以在课堂上使用的任务。然而，为了准备考试，教师通常需要修改教科书中的任务，以测试学生的问题解决技能和概念理解，同时避免依赖对结果的记忆。为了考虑到这一实践，并确保两组之间进行公平比较，研究人员要求对照组参与者对教科书任务进行重大修改（修改内容不能仅限于更改给定数值）。此外，干预组参与者无法使用物理教科书，而对照组无法使用 ChatGPT。

在创建任务之前，干预组学生被要求将给定的三个提示词输入 ChatGPT（1.“围绕歌德《浮士德》这本书创建一个问题。”；2.“围绕赫尔曼·黑塞的《德米安》一书中德米安的作用，为高中 11 年级学生创建一个问题。”；3.“创建一道选择题，考查《哈利·波特与魔法石》中魁地奇对哈利·波特的重要性。”），以熟悉 ChatGPT 在回应方面的敏感度和灵活性。完成这些提示词后，学生有机会尝试自行输入内容，然后再进入概念任务设计阶段。

我们要求各组参与者提交自己创作或改编的问题，以及启发其进行任务开发或改编的原始问题（如有），原始问题可来自教科书或 ChatGPT。这样做有两个目的：首先，确保对照组的参与者不会直接照抄教科书中的问题而不作修-改。若出现这种情况，研究结果便不具效度，因为我们的目标是评估参与者创作原创评估任务的能力。其次，我们旨在观察参与者对 ChatGTP 提出的任务作了多大程度的修改。这使我们能够评估 ChatGPT 生成高质量评估任务的能力，以及教育工作者使用该工具时所需的总工作量。

我们已保存聊天记录，以核实学生的输入。

由于教育工作者创造合适评估的能力受到其自身物理学概念理解的影响，研究使用半长度版力概念测验（FCI）2 [59]评估参与者的先前知识。之所以选择该测试，是因为它涵盖了学生据以设计任务的上述概念。充分理解学科内容对于弥补 Chat-GTP 在生成一致任务方面的潜在局限性可能至关重要 [54, 55]。需要指出的是，标准教科书通常不受这一局限影响，因为它们会定期接受质量检查和修订。不过，即使如此，教育工作者在修改教科书任务以用于期末考试时，也必须充分掌握所评估的概念和任务质量。

最后，鉴于大型语言模型是一种相对较新的工具，可能会带来与用户相关的挑战，因此我们对干预组参与者进行了可用性调查。调查包括系统可用性量表（SUS）中的十个问题 [60]，以及技术接受模型2（TAM2）中关于感知有用性和输出质量的六个问题 [61]。此外，还增加了一个问题，以评估 ChatGPT 平均响应时间意外变慢是否影响用户体验（见补充材料）。每个问题均采用五点李克特量表作答：完全不同意、不同意、中立、同意、完全同意。研究结束时，我们请参与者提供其人口统计数据。

C. 已开发任务的定性数据分析

参与者共开发了 $N = 103$ 项评估任务（干预组$N = 51$ 项，对照组 $N = 52$ 项）。根据既有文献，我们确定了八个类别，用于评估已开发的评估任务：具体性（量表： 0=不适用；1=适用）、清晰度（量表：0/1）、正确性（量表：0/1）、是否具有误导性（量表：0/1）、难度适宜性（量表：0/1）、情境（量表：0/1）、与目标概念的相关性（五点李克特量表：完全不相关、较不相关、不相关与相关方面数量大致相同、较相关、完全相关），以及总体质量（五点李克特量表：非常低、较低、中等、较高、非常高）（有关这些类别的说明，请参阅补充材料）。

两名拥有六年以上经验的独立评定者依据以下八个类别对每项任务进行评分。他们的经验包括教授未来物理教师、开展物理教育研究，以及为高中和大学物理编制概念性问题。对于前六个类别，计算了两名评定者之间的 Cohen’s κ 一致性（见表 I）；其余两个类别则取平均值（见结果）。

根据Landis和Koch的说法，这些Cohen’s值

类别	科恩卡帕系数
具体性	0.43
清晰度	0.52
正确性	0.56
不具误导性	0.35
难度适宜性	0.76
情境	0.66
平均值	0.55

表 I. 所开发任务评分的评定者间信度。

κ 可以解读为：对于任务措辞是否具有误导性这一项，最低情况下的一致性尚可（κ =0.35）；对于任务难度是否适合10^{th} 高中物理课的年级水平 [62]这一项，则达到较高的一致性（$\kappa = 0.76$）。总体而言，两位评分者之间的一致性中等（κ =0.55）。评分意见不一致之处通过讨论解决

IV. 结果

A. 所生成任务的质量

解决分歧后，我们比较了两组在八个类别上的差异（见图1）。 t检验显示，两组在任务具体性、正确性、措辞是否具有误导性以及任务难度是否适宜方面均无显著差异（各项 $p > 0.05$）。值得注意的是，两组在目标组任务适切性方面都取得了（近乎）满分，任务正确性也都达到了很高水平。相比之下，两组的任务具体性得分都较低，约为0.4，表明两组学生都难以为任务提供足够的信息，使其能够得以解答。此外，两组在任务清晰度（p=0.04）方面存在显著差异，效应量较小（Cohen's $d = 0.38$）；在任务情境（$p = 6 \cdot 10^{-5}$）方面也存在显著差异，效应量较大（ $d = 1.04$ ）。这两项指标中，使用教科书的学生得分都更高。

表 II 展示了学生在 ChatGPT 组和教材组中设计的一些任务示例。在第一个示例中，两个组的任务在八个类别的评分中均达到了总体较高的质量，例如，任务内容正确、清晰、简洁，嵌入了有意义的情境，并且对应目标概念。在第二个示例中，ChatGPT 组的任务遗漏了汽车从静止状态加速到 100 km/h 所需的时间，因此说明不够充分，无法确定汽车的加速度。同样，在教材组的第二个任务中，参与者提到没有发生摩擦力，但未能说明是否有其他力作用于物体。在第三个示例中，两个组的任务都缺乏情境。

此外，我们还考察了得分是否

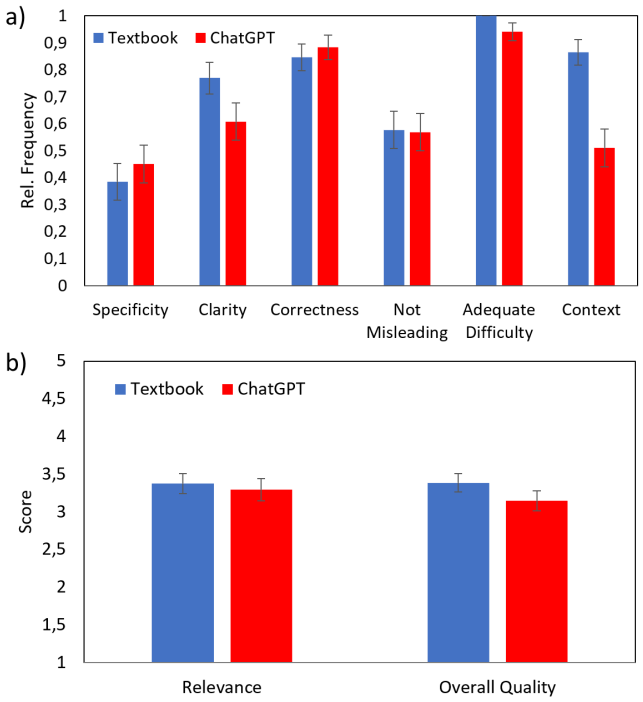


图 1. 对照组（教材组）与干预组（ChatGPT 组）之间的差异：（a）在 0（不适用）到 1（适用）的量表上评定的六个类别；（b）在 1（非常低）到 5（非常高）的李克特量表上评定的类别。

与各组学生的先前知识相关。干预组参与者的力概念测验平均得分为 0.62 ± 0.05，而对照组参与者的力概念测验平均得分为0.69 ± 0.05 （总体平均分： 0.65 ± 0.05），但两组之间的差异并不显著（$p > 0.05$）。为此，我们对学生的力概念测验得分与其在前六个类别中所创设任务的评分进行了逻辑回归分析，这些任务的评分采用 0 和 1 两个等级。我们发现，力概念测验得分与前六个类别的评分之间不存在显著相关关系。此外，我们对学生的力概念测验得分与最后两个类别的评分进行了线性回归分析，这两个类别的评分采用 1 至 5 的等级。我们同样未发现这两项指标之间存在线性关系。因此，在本研究中，先前知识对所创设任务的质量没有显著影响。

B. 对教科书和 ChatGPT生成的任务的调整

除了评估八个类别中的任务质量外，我们还评估了学生相较于ChatGPT生成的任务对教科书任务所做的修改。我们评估了这些修改是否导致任务质量提高（+1）或降低（-1）。总体而言，ChatGPT 组中有一名参与者在一项任务中没有使用Chat

方面	ChatGPT 3.5	教科书
总体质量较高	一辆汽车从 20 m/s 的初速度开始加速 在 10 秒内达到 60 m/s 的末速度- 秒。计算这段时间内汽车的加速度 假设加速度恒定, 。	说明为什么地球轨道上的太空碎片可能非常 危险，会对国际空间站这样的空间站构成威胁。请运用 牛顿第一定律进行论证。
缺乏具体性	一辆汽车从静止状态加速到 100 km/h。 1. 解释物理量加速度与速度之间的区别， 并说明它们的单位。 并解释其单位。 2. 计算汽车的加速度，单位为 m/s^2。	一个物体以速度 v_0 做无摩擦运动。 10 秒后物体的速度是多少？请说明理由 并用牛顿定律解释你的判断。
缺少情境	牛顿第一定律如何描述.....的行为 根据物体的运动或静止状态?	从静止状态开始做匀加速运动时，平均速度如何与 终端速度相关? 你可以使用哪些工具 来完成这项任务?

表 II. 使用 ChatGPT 3.5 或教科书生成的、体现高质量、缺乏具体性，或缺少情境。

GPT作为支持工具（占该组所有任务的2%），而是在没有任何支持的情况下自行创建了该任务。相比之下，教材组的参与者在没有使用教科书作为资源的情况下完成了42项任务（占该组所有任务的81%）。因此，教材组的参与者使用教科书完成了10项任务（占该组所有任务的19%），并对给定的问题进行了调整。相比之下，ChatGPT 组的参与者使用ChatGPT完成了50项任务（占该组所有任务的98%），并在12种情况下对ChatGPT任务进行了调整（占该组所有任务的24%）。这意味着ChatGPT 组的学生直接采用了ChatGPT提供的38项任务（占该组所有任务的75%）。

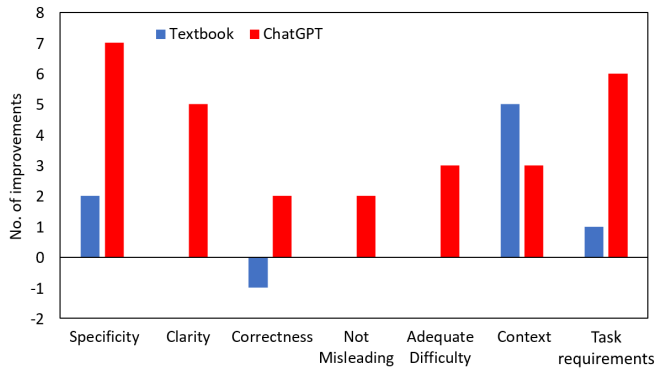


图 2. 学生对教材任务或ChatGPT生成的任务在七个类别中所做改进的数量。负值表示参与者的改编降低了任务质量。

图 2 展示了学生在七个类别中所做改进的数量。正值表示更多修改使任务得到改进，而负值表示更多修改导致任务质量下降。前六个类别与质量评分中使用的类别相同，但省略了有关与概念图相关性的类别。总体质量也被省略，并新增了一个反映任务要求变化的类别。在这一新类别中，我们评估了未来物理教师是否调整了任务类型，或调整了任务要求学生完成的内容。表 III 列出了参与者使用 ChatGPT生成的任务或从教科书中提取的任务所做的四项此类修改。

表 III 中的示例 1 展示了 ChatGPT 组的一名参与者对任务类型所做的调整。在这个例子中，这位准物理教师未能意识到，根据现有信息，任务第 2 部分中航天器的速度无法确定。尽管如此，参与者仍将重点转向修改任务第 3 部分的要求，将定量任务改为描述性任务。图 2 显示，与教材组的参与者相比，使用 ChatGPT 的参与者做出的改进更多。出现这一结果是意料之中的，因为教材组的参与者主要独立于教科书开发任务，因此没有记录到对教科书任务的修改。相比之下，ChatGPT 组的学生主要调整了任务具体性和要求。例如，表 III 中的示例 2 展示了 ChatGPT 组的一名参与者在 ChatGPT 生成任务后对任务具体性所做的修改。在这个例子中，由于不明确汽车的加速度是否恒定，因此无法确定其加速度。参与者认识到了这一局限，随后

示例编号	组别	ChatGPT生成的任务或教科书任务	参与者改编的任务
1	ChatGPT	一艘宇宙飞船的质量为 2000 千克，正处于绕 X行星 运行的轨道上。引力 X行星对航天器的力为 20000 N。 1. 计算太空的加速度- 针对 X 行星进行设计。 2. 航天器的速度是多少? 轨道? 3. 该物体的轨迹有多大? 航天器的速度发生变化时 器会怎样改变?	一艘宇宙飞船的质量为 2000 kg，正处于绕 X行星 运行的轨道上。X行星的引力对航天器的作用力为 20000 N。 1. 计算航天器的加速度 相对于X行星的航天器。 2. 航天器在 轨道? 3. 描述航天器的轨 迹（当航天器的速度 发生变化时）。
2	ChatGPT	一辆汽车的初速度为 20 m/s，随后加速在 10 秒内达到 60 m/s 的末速度。计-算这辆汽车的加速度。	一辆汽车的初速度为 20 m/s，随后加速在 10 秒内达到 60 m/s 的末速度。计-计算汽车在这段时间内的加速度 假设加速度保持恒定。
3	教科书	火箭推进的原因 与你放飞的充气气球相同，是 相互作用原理。解释这一说法。	为了让一架波音747获得足够的推力 飞行，牛顿相互作用定律的作用方式很像放开一个充满空气的气球。请解释 这一原理在波音747飞行情境中的应用（或起飞时的应用）。
4	教科书	塞尔玛（m1=60kg）助跑后跳上 一辆静止的雪橇（m2=12kg）。随后两者以速度 u=2.5 m/s 继续前进。 1. 计算塞尔玛落到雪橇上时的速 她落在雪橇上的位置。 2. 证明动能并未守 在此过程中提供。	塞尔玛（m1=60kg）助跑跳上 一辆正在移动的雪橇（m2=12kg，v2=2 m/s）。两者继续一起运动。 1. 假设没有摩擦力且没有坡度，计算 它们共同的最终速度。 2. 计算碰撞时的总动量 。

表 III. 参与者对 ChatGPT 编写或从教科书中提取的任务所做修改示例。

据此修改了任务，最终形成了定义明确且具体的任务。

教材组主要侧重于修改任务的情境。表 III 中的示例 3 表明，两个任务都是正确的，情境仅有轻微变化。然而，也有参与者错误修改教科书任务的情况。例如，表 III 中的示例 4 表明，参与者未能注意到该任务关注的并非速度与加速度之间的关系，而是非弹性碰撞中的动量守恒。此外，参与者也没有纳入塞尔玛的初速度。因此，任务第 1) 部分的末速度和第 2) 部分的总动量都无法再确定。

C. ChatGPT 3.5 创建物理任务的可用性

为了评估参与者如何看待他们从 ChatGPT 获得的输出，以及他们认为重新.....有多容易或困难在收到系统提供的质量足够高的回复后，我们请参与者评价系统可用性、感知有用性和输出质量（图 IV C）。参与者对系统可用性的平均评分为 4.1，表明他们认同该系统易于使用、各项功能整合良好，并且有信心使用该系统。关于开发物理任务时的感知有用性，参与者的评价为中立（平均值为 3.24）。这意味着他们既没有觉得工作效率或成效有所提升，也没有觉得开发物理任务的过程因此变慢。该量表还包括一个问题，询问等待 ChatGPT 回复的时间是否使系统使用变得复杂。平均而言，学生并不认同（平均值为 2.09）回复时间使系统使用变得复杂。此外，参与者对 ChatGPT 输出质量的评价为中立，但平均值最低，仅为 2.76，这表明参与者往往难以处理系统的输出。

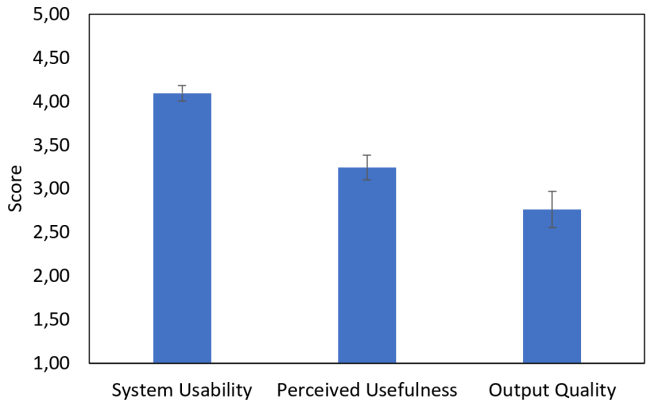


图 3. 参与者对 ChatGPT 的可用性、感知有用性和输出质量的评分，采用五点李克特量表： 1=非常不同意， 2= 不同意， 3= 中立， 4=同意， 5=非常同意。分析时，已反向处理措辞为否定的条目，因此高分反映学生的正面评价。

V. 讨论

本文旨在探讨三个研究问题。第一个研究问题关注参与者使用 ChatGPT 生成的任务，与学生使用教科书编写的任务相比，质量如何。两组的任务正确性都非常高，超过了概念测验的平均分。假设生成正确评估任务的难度与半版力概念测验的平均难度相当，那么这一发现表明，ChatGPT 能够弥补未来物理教师可能存在的概念理解困难，从而降低这些困难转化为评估任务的可能性。

此外，值得注意的是，两组的任务具体性都较低，这表明完成任务所需的信息有所缺失。无论使用 ChatGPT 3.5 还是教科书，都无法弥补未来物理教师在这方面的不足。未来版本的大型语言模型能否克服这一缺陷，或者能否通过面向未来物理教师的学习干预或其他支持方式加以解决，值得进一步探讨。此外，参与者使用 ChatGPT 创建的任务显著较少地置于情境之中。不过，如果参与者明确提示，ChatGPT 生成的一些任务也包含情境。因此，ChatGPT 组得分较低并非系统所致，而是因为学生没有意识到情境对于面向高中十年级学生的物理评估任务的价值。教材组的学生在这一类别中得分较高，无论他们是修改给定的教科书任务，还是在没有教科书任务启发的情况下设计物理任务。

第二个研究问题关注的是数量以及学生对 ChatGPT 生成的任务和教材任务所做的修改类型。值得注意的是，教材组中 81% 的任务并非基于给定的教材任务改编，而是学生未使用该资源自行设计的。这一观察表明，对该组学生而言，凭记忆或自行设计任务，比查阅教科书中的任务更省力。这种行为可能取决于学生对教科书的熟悉程度，以及他们此前已在相同情境中设计了多少评估任务并将自己的想法融入其中。结果显示，一些参与者意识到 ChatGPT 创建的任务具体性不足，并对其进行了改进，但大多数参与者没有意识到这一点。

此外，结果表明，部分学生修改现有教科书中的任务后，导致任务出现错误。尽管这种情况仅出现在三项任务中（占该组所有任务的 6%），且该组所有任务的正确性与 ChatGPT 组相当，但在将给定任务布置给学生之前，额外使用大型语言模型来解答该任务，或许可以弥补这一不足。

第三个研究问题探讨了 ChatGPT 输出的可用性、感知有用性和质量。我们发现，参与者对 ChatGPT 3.5 的可用性评价较高，但认为 ChatGPT 生成的物理任务质量略低于中性水平。未来版本的 ChatGPT 是否能提升输出质量评价，以及如果学生接受了如何编写提示词以让 ChatGPT 生成物理任务的培训，这种评价是否会有所改变，都值得进一步探讨。

VI. 结论

本研究评估了未来物理教师使用 ChatGPT 3.5 开发的任务质量，并将其与使用教科书开发的任务进行比较。ChatGPT 所创建任务的正确性、难度适宜任务的比例及总体质量，与能够使用教科书的参与者所创建的任务相当。对于大型语言模型而言，这是一个显著成就，因为目前使用教科书开发物理评估任务是最先进水平，并要求教师具备评估知识、概念知识和问题解决技能 [63]。与此同时，大型语言模型领域目前正在迅速发展，它们很可能很快就能更熟练地完成任务开发等学科专属练习。

我们发现，任务中的一些不足之处（如具体性）无法通过教科书或 ChatGPT 3.5 来弥补；而其他方面则需要 ChatGPT 用户明确提示，例如任务的情境。

总之，我们证明了 ChatGPT3.5 等先进的大型语言模型能够有效支持物理教师的日常实践，即

开发评估任务。在教育领域为物理教师提供专项培训将大有裨益。

用于设计提示词、以有效使用大型语言模型的程序。

[1] N. AFT, 《教育测量：问题与实践》9, 30. [2] J. Gess-Newsome, J. A. Taylor, J. Carlson, A. L. Gardner, C. D. Wilson, and M. A. Stuhlsatz, International Journal of Science Education 41, 944 (2019年). [3] P. W. Airasian, *Classroom assessment: Concepts and applications* 2001 (ERIC,).

[4] C. A. Mertler, (1998). [5] R. J. Stiggins, High School Magazine 6, 20 (1999). [6] W. J. Popham, 《教育测量：问题与实践》22, 45 (2003年). [7] L. Eyal, Journal of Educational Technology & Society 15, 37 2012( ). [8] E. Kasneci, K. Seßler, S. K¨uchemann, M. Bannert, D. De- mentieva, F. Fischer, U. Gasser, G. Groh, S. G¨unnemann, E. H¨ullermeier, *et al.*, Learning and Individual Differences 103, 102274 (2023年). [9] H. E. Fischer and A. Kauertz, in *Physics Education* (Springer, 2022年) pp. 231–267. [10] J. Leisen, Mathematische und Naturwissenschaftliche Un- terricht 54, 401 (2001). [11] K. Neumann, T. Viering, W. J. Boone, and H. E. Fischer, Journal of research in science teaching 50, 162 (2013年). [12] J. F. Osborne, J. B. Henderson, A. MacPherson, E. Szu, A. Wild, and S.-Y. Yao, Journal of research in science teaching 53, 821 (2016). [13] V. Villarroel, D. Boud, S. Bloxham, D. Bruna, and C. Bruna, Innovations in Education and Teaching In- ternational 57, 38 (2020年). [14] P. Sinaga and S. Feranie, International Journal of Instruc- tion 10, 69 (2017). [15] J. M. Calmer, Pedagogical Research 4, em0041 (2019年). [16] S. der St¨andigen Konferenz der Kultusminister der L¨ander in der Bundesrepublik Deutschland, Bildungsstandards im fach physik f¨ur den mittleren schulabschluss. beschluss vom 16.12. 2004 (2005). [17] K. Kremer, H. E. Fischer, A. Kauertz, J. Mayer, E. Sumfleth, and M. Walpuski, Making it tangible: science education中的学习成果, 201 (2012). [18] J. Leisen, Mathematische und Naturwissenschaftliche Un- terricht 59, 260 (2006). [19] N. Akben, Research in Science Education 50, 1143 (2020年). [20] T. Rowland, P. Huckstep, and A. Th waites, Journal of Mathematics Teacher Education 8, 255 (2005). [21] L. S. Shulman, in *Handbook of research on teaching*, edited by M. C. Wittrock (Macmillan, 1986) pp. 3–36+. [22] J. Gess-Newsome, J. Taylor, J. Carlson, A. Gardner, C. Wilson, and M. Stuhlsatz, International Journal of Science Education 41 (2017). [23] T. Kleickmann, D. Richter, M. Kunter, J. Elsner, M. Besser, S. Krauss, and J. Baumert, Journal of Teacher Education 64, 90 (2012). [24] L. S. Shulman, Harvard educational review 57, 1 (1987). [25] R. A. Edwards, R. Hyde, M. C. O'Connor, and J. Oldham (2015). [26] Y. Lee, R. Capraro, and M. Capraro, International Electronic Journal of Mathematics Education 13, 10.12973/iejme/2698（2018年）。[27] E. Silver, 《For the Learning of Mathematics》14，19（1994年）。[28] C. Bonotto, 《Educational Studies in Mathematics》83，37（2013年）。[29] J. Cai 和 R. Leikin, 《Educational Studies in Mathematics》105，287（2020年）。[30] B. Koichu、G. Harel 和 A. Manaster, 《Instructional Science》41，681（2013年）。[31] T. Lowrie, 《Contemporary Issues in Early Childhood》3，354（2002年）。[32] S. Crespo, 《Educational Studies in Mathematics》52，243（2003年）。[33] S. Crespo 和 N. Sinclair, 《Journal of Mathematics Teacher Education》11，395（2008年）。[34] J. Cai 和 S. Hwang, 《The Journal of Mathematical Behavior》21，401（2002年）。[35] E. A. Silver 和 J. Cai, 《Journal for Research in Mathematics Education》27，521（1996年）。[36] J. Cai 和 S. Hwang, 《International Journal of Educational Research》102，101391（2020年）。[37] OpenAI，OpenAI Chat（2023年），访问日期：2023年3月28日。[38] L. Floridi 和 M. Chiriatti, 《Minds and Machines》30，681（2020年）。[39] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，arXiv 预印本 arXiv:1810.04805（2018年）。[40] Y. Liu、M. Ott、N. Goyal、J. Du、M. Joshi、D. Chen、O. Levy、M. Lewis、L. Zettlemoyer 和 V. Stoyanov，arXiv 预印本 arXiv:1907.11692（2019年）。[41] M. Zhu、O. L. Liu 和 H.-S. Lee, 《Computers & Education》143，103668（2020年）。[42] J. P. Bernius、S. Krusche 和 B. Bruegge, 《Computers and Education: Artificial Intelligence》3，100081（2022年）。[43] E. Kasneci、K. Sessler、S. K¨uchemann、M. Bannert、D. De- mentieva、F. Fischer、U. Gasser、G. Groh、S. G¨unnemann、E. H¨ullermeier、S. Krusche、G. Kutyniok、T. Michaeli、C. Nerdel、J. Pfeffer、O. Poquet、M. Sailer、A. Schmidt、T. Seidel、M. Stadler、J. Weller、J. Kuhn 和 G. Kasneci, 《Learning and Individual Differences》103，102274（2023年）。[44] V. Raina 和 M. Gales，arXiv 预印本 arXiv:2209.11830（2022年）。[45] R. Abdelghani、Y.-H. Wang、X. Yuan、T. Wang、H. Sauz´eon 和 P.-Y. Oudeyer，arXiv 预印本 arXiv:2211.14228（2022年）。[46] S. Sarsa、P. Denny、A. Hellas 和 J. Leinonen，载于 *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1* 2022 27 （），第43页。[47] Q. Jia、J. Cui、Y. Xiao、C. Liu、P. Rashid 和 E. F. Gehringer，arXiv 预印本 arXiv:2110.03895（2021年）。[48] J. Jeon, 《Computer Assisted Language Learning》，1（2021年）。[49] M. Bao, 《Arab World English Journal》（AWEJ）CALL 专刊（2019年）。[50] T.-Y. Tai 和 H. H.-J. Chen, 《Interactive Learning Environments》，1（2020年）。[51] S. Polak、G. Schiavo 和 M. Zancanaro，载于 *CHI Conference on Human Factors in Computing Systems Extended Abstracts* 2022 1 7 （），第页。

[52] R. Chocarro、M. Cortinas 和 G. Marcos-Mat´as, 《Educational Studies》**49**，295（2023年）。[53] W. Holmes、M. Bialik 和 C. Fadel, 《教育人工智能》（2020年）。[54] ChatGPT, 《ChatGPT 的局限性》，与 ChatGPT 的在线对话（2023年），来源：URL（如适用）。[55] ChatGPT, 《基于 GPT-4 和 GPT-3.5 的 ChatGPT 版本的局限性》，与 ChatGPT 的在线对话（2023年），来源：URL（如适用）。[56]OpenAI, 《GPT-4：高级人工智能语言模型》（2023年），访问日期：2023年3月28日。[57] OpenAI, 《GPT-4 技术报告》（2023年），arXiv:2303.08774 [cs.CL]。[58] R. Dietrich 和 F. Finkenberg, *Physik 10 Gymnasium*，(*Bayern* 2022 C.C. Buchner，)。[59] J. Han、L. Bao、L. Chen、T. Cai、Y. Pi、S. Zhou、Y. Tu 和 K. Koenig, 《物理教育研究专题评论》**11**，010112（2015）。[60] M. Schaefer、B. Xu、H. Flor 和 L. G. Cohen, 《人脑图谱》**30**，2722（2009）。[61] V. Venkatesh 和 F. D. Davis, 《管理科学》 **46**，186 （ 2000 ）。[62] J. R. Landis 和 G. G. Koch, 《生物统计学》，159（1977）。[63] S. M. Brookhart, 《教育测量：问题与实践》**30**，3（2011）。