小组评价与 ChatGPT-4o 评价的质量比较研究

A Comparative Study on the Quality of Group Assessment and ChatGPT-4o Assessment

刘婷玉[1]，张悦[1]，杨现民[1]，李新[1*]
[1]江苏师范大学智慧教育学院
[*] lixin_407@163.com

【摘要】 在全球教育改革与技术迅猛发展的趋势下，高等教育评价体系正经历深刻变革。同伴互评与人工智能评价工具成为研究重点。然而，对小组评价与 ChatGPT 评价的比较研究仍然不足，尤其是 ChatGPT-4o 的出现为教育评价领域注入活力。本研究采用 Mann-Whitney U 检验和访谈主题分析法，系统比较了小组和 ChatGPT-4o 评价质量的差异，并探究学生对二者的接受程度。结果显示，第一轮评价中二者质量无显著差异，第二轮中二者在建设性、准确性、具体性维度存在显著差异。访谈显示，学生对两种评价看法不一，二者具有互补作用。研究为教师合理选用评价方式提供参考，助力教育评价向多元化、智能化发展。
【关键词】 小组评价；ChatGPT-4o 评价；高等教育；智慧学习环境设计

Abstract: *With the rapid development of global education reform and technology, the evaluation system of higher education is undergoing profound changes. Peer review and artificial intelligence evaluation tools have become the focus of research. However, the comparative study of group evaluation and ChatGPT is still insufficient, especially the emergence of ChatGPT-4o has injected vitality into the field of educational evaluation. In this study, Mann-Whitney U test and interview topic analysis were used to systematically compare the differences in evaluation quality between the group and ChatGPT-4o, and to explore students ' acceptance of the two. The results show that there is no significant difference in the quality of the two in the first round of evaluation, and there are significant differences in the dimensions of constructiveness, accuracy and specificity in the second round. The interview shows that students have different views on the two evaluations, and the two have complementary effects. The research provides a reference for teachers to choose evaluation methods reasonably, and helps education evaluation develop towards diversification and intelligence.*
Keywords: Group assessment, ChatGPT-4o assessment, higher education, smart learning environment design

1. 问题提出

在全球教育改革与技术飞速发展的双重驱动下，高等教育领域的评价体系正经历深刻变革。评价作为学生学习过程中的核心环节，不仅关乎学习成效的监测，还直接影响学生能力的培养。在这一背景下，同伴互评逐渐成为教育实践中的重要评估机制，鉴于其能够有效促进学生的批判性思维发展(Noroozi et al., 2016)、增强协作学习效果及提升学习自主性(Villamil & De Guerrero, 1998)。然而，传统一对一的同伴互评模式由于学生认知局限和主观偏差等因素，往往存在评价质量不稳定的问题。小组评价作为一种集体参与的互评形式，凭借其对团队协作和集体智慧的充分调动，不仅能强化个体与团队的学习成效，还能显著提高协作学习的动力(Katz et al., 2023)。因此，小组评价在学术与专业领域的多种情境中得到了广泛应用。此外，随着教育技术的飞速发展，在技术辅助评估领域，人工智能（AI）技术的引入，为评价机制的创新与优化提供了全新可能性。近年来，基于 AI 的评价工具，如 ChatGPT，其文本分析与反馈能力已成为研究热点(Lu et al., 2024)。然而，现有研究多聚焦于 ChatGPT 与教师评价的比较分析，对小组评价与 ChatGPT 评价之间的异同探讨则相对缺乏。且由于技术局限，早期版本的 ChatGPT 主要适用于文本内容的评估，难以应用于复杂的多模态评价场景。随着 ChatGPT-4o 的问世，其在处理多模态内容方面展现出显著潜力，这不仅为扩展其应用场景带

来了新的可能，也为提升评估效果提供了技术支持。这一技术突破引发了学术界对其在多模态评估环境中表现的深入探讨。因此，本研究旨在比较小组评价与 ChatGPT-4o 评价的效果，分析二者在多模态评价任务中的表现差异，从而为教师在高等教育评价中使用二者提供参考。

基于此，本研究的研究问题如下：

（1）小组生成的评价与 ChatGPT-4o 生成的评价在质量上存在哪些异同？

（2）学生对小组评价与 ChatGPT-4o 评价的看法如何？

2. 文献综述

2.1. 同伴互评

同伴互评作为一种重要的教学策略，要求学习者对同伴的作品进行全面评价，包括作品的质量、价值及其学习的总体成效(Topping, 1998)。相关实证研究表明，同伴互评不仅能够有效提升学生的批判性思维能力(Joordens et al., 2009; Topping, 2009; Wang et al., 2017)，还在高阶思维技能的培养(Topping, 2017)、社会能力的提升(Ching & Hsu, 2016)、学习动机的增强(Hsia et al., 2016)以及整体学术表现的改进(Zheng et al., 2016)方面具有显著作用。正因如此，同伴互评已成为推动学习者主动参与知识建构的重要手段。从评价的组织形式上看，同伴互评可以分为一对一式、小组成员式和小组之间式的相互评价。其中，小组间的相互评价，是将学生群体的状态、行为以及成果作为评价对象开展的一项活动（孙明霞等，2024）。作为一种依赖个体主观评价的方式，一对一式同伴互评的可靠性和有效性仍然存在争议(Topping, 2009)，研究指出，评价过程中的主观偏差和人际关系因素，可能影响评价质量，甚至引发学生之间的人际关系紧张(Panadero, 2016)。为缓解上述问题，高等教育课堂中逐渐倾向于同伴评价的另一种形式——小组评价。小组评价是指一个小组对另一小组的知识掌握和技能运用情况进行集体评价(Zhang et al., 2023)。通过集体讨论与决策，小组评价有望减少个体主观性对评价结果的干扰。研究发现，小组评价相较于个体评价在多方面展现出独特优势，如增强学生之间的互动性(Tan & Chen, 2022)、进一步促进批判性思维的培养(Chen et al., 2021)。例如，Guo 等(2022)在研究小组合作对作文反馈的影响时发现，学生在小组评价中更倾向于关注与内容相关的问题，并能够提供更为批判性和建设性的反馈。尽管小组评价在实践中展现出诸多潜力，但相关研究目前仍显不足，对于小组评价的质量、学生对这一评估形式的接受度等问题，尚缺乏系统性探讨，这为如何进一步优化这一评价方式提供了亟待解决的重要议题。

2.2. ChatGPT 评价

人工智能（AI）技术在教育领域的广泛应用正推动教育模式的深刻变革。其中，ChatGPT 的出现被认为是教育技术发展的重要里程碑，展现出优化教学与评估模式的潜力。作为人工智能领域的一项前沿成果，其在跨领域的自然语言处理方面展现出卓越的能力。该技术能够以类似人类的对话形式与用户进行互动，并且能够根据用户的提问，生成具有高质量的答案（苗逢春，2023），具备具备激发思考的内容创造能力、对话场景的理解力、有序任务的执行技巧以及编程语言的解读能力（李毅等，2024）。其即时性、灵活性和积极的语言风格，使学习者能够获得更加个性化的学习支持，从而激发学习者的学习动机(Guo & Wang, 2024)。Lu 等人(2024)发现，ChatGPT 生成的反馈在数量和总结性评论方面具有优势，有助于学习者的反思与改进(Lu et al., 2024; Barrot, 2023)。2024 年 5 月，OpenAI 发布了新一代 ChatGPT-4o 版本。这一版本在速度、功能和语言生成能力上实现了显著突破，不仅能够快速生成与任务要求高度相关的精准反馈，还通过幽默的语言风格增强了交互体验，被视为自然语言处理技术的一次飞跃(Kleinman, 2024)。此外，ChatGPT-4o 集成了多模态信息处理技术，能够识别和处理包括文本与图像在内的多种数据形式(Wiggers, 2024)，为解决复杂的开放性任务提供了技术支持，进一步拓展了其在高等教育中的应用场景。尽管如此，目前关于 ChatGPT 在开放性任务评估中实际应用效果的研究仍显不足，尤其是在其与小组评价评估质量差异的比较上，尚缺乏系统且深入的研究。对于高校教师而言，全面了解 ChatGPT 与小组评价在评估效果上的不同表现，不仅能够帮助他们更科学地选择评估方式，还能为提升教学效果提供有力支持。

3. 研究设计

3.1. 研究对象

研究在江苏某高校教育技术学专业开设的名为“智慧学习环境设计”的课堂中进行，参与者为 39 名本科生，其中男生 13 人、女生 26 人，这些学生被随机分配到 12 个小组，每组由 3-4 人组成。

3.2. 研究工具

3.2.1. 方案评价量规

小组智慧学习环境设计方案从五个方面进行评价：创新性、实用性、技术实现性、用户友好性和环境适应性，每个维度的评分等级如下：优秀（21-25 分）、良好（16-20 分）、一般（11-15 分）、差（6-10 分）和非常差（1-5 分）。我们为每个等级的水平都提供了详细的描述以供小组和 ChatGPT-4o 对设计方案进行评分并阐述理由。该评分维度及标准由在该领域经验丰富的讲师和助教团队细致商讨后最终确定。为了确保小组分数的准确性，一位讲师和两名助教同时评定了所有的小组方案,且两名助教审查了所有的小组评分，发现小组总体上遵循了标准，并为他们的评分提供了实际理由。

3.2.2. 学生评分问卷

学生从有效性、建设性、准确性、完整性和具体性(Guo & Wang, 2024; Banihashem et al., 2024; Steiss et al., 2024)五个维度来评价小组反馈和 ChatGPT-4o 反馈的质量。学生按照 1 到 5 的等级对每个维度进行评分，1 表示最低，5 表示最高。对于每个维度，学生还需要阐明他们给出该分数的理由。

3.3. 研究过程

本研究共进行两轮，每轮两周，两轮遵循相同的实验流程，如图 1 所示。在实验进行之前学生已经通过课程学习积累了智慧学习环境设计的相关知识。在实验的第一周，各小组根据讲师要求起草智慧学习环境设计方案（以下简称“方案”）。在第二周，每组的方案由互评小组和 ChatGPT-4o 根据给定量规进行评价，评价包括评分和给出评分背后的理由。随后，小组评价和 ChatGPT-4o 评价以“评价 1”和“评价 2”的形式盲法呈现给各小组，在各小组针对两份评价进行深入的讨论后，由每位学生对二者进行评分，并说明评分背后的原因。两轮实验结束后，我们从高、中、低三个得分水平中随机抽取两个小组参加访谈，访谈包括：回忆两轮实验中评价 1 和评价 2 的内容，并分别谈论对两种评价的看法。

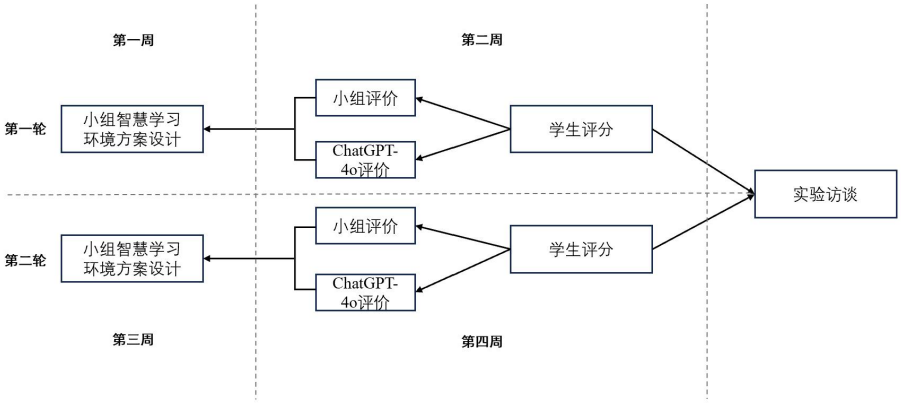

图 1 实验流程

3.4. 数据收集与分析

本研究数据收集分为两个阶段。第一阶段，我们利用问卷星平台发放问卷，以获取学生在五个维度上对两种反馈方式的评分数据，并要求学生在评分时说明理由。在第二阶段，我们收集了最终的访谈数据。

为了探讨两种评价方式的质量差异，我们对学生给两者的评分数据进行了深入分析。首先，我们比较了两轮实验中两种评价方式在各维度上的得分均值，以初步探索二者的差异；随后，

采用 Mann-Whitney U 检验，进一步探索两者在五个维度上的评价质量是否存在显著差异。为了探讨学生对小组评价与 ChatGPT-4o 评价的看法，我们对访谈数据进行了主题分析。依据 Braun 和 Clarke（2006）的主题分析框架，研究人员通过系统的编码过程展开分析。首先，研究人员独立阅读访谈内容并标记初步代码，这些代码随后被归类为积极和消极的看法。通过迭代分析，反复出现的主题被识别、概括并命名。为确保研究方法的严谨性，我们与领域专家进行了同行汇报，以进一步完善编码方案。

4. 研究结果与讨论

4.1. 小组评价和 ChatGPT-4o 评价的质量差异

表 1 展示了两轮实验中，小组评价与 ChatGPT-4o 评价在有效性、建设性、准确性、完整性和具体性五个维度上的统计分析结果。在第一轮实验中，小组评价与 ChatGPT-4o 评价的均值差异较小，除具体性外，小组在其他各维度的平均得分均高于 ChatGPT-4o。然而，在第二轮实验中，两者的均值差异扩大，ChatGPT-4o 在所有维度上的平均得分均高于小组评价。

表 1 两轮实验的统计分析结果

质量维度	第一轮				第二轮			
	小组评价		ChatGPT-4o 评价		小组评价		ChatGPT-4o 评价	
	均值	标准差	均值	标准差	均值	标准差	均值	标准差
有效性	3.79	0.732	3.69	0.766	3.77	0.667	4.05	0.647
建设性	3.77	0.842	3.67	0.772	3.64	0.537	4.10	0.598
准确性	3.90	0.680	3.54	0.854	3.74	0.637	4.13	0.615
完整性	3.72	0.686	3.51	0.914	3.77	0.667	3.92	0.664
具体性	3.54	0.854	3.74	0.818	3.59	0.715	4.00	0.688

我们采用 Mann-Whitney U 检验进一步探讨了两轮实验中小组评价与 ChatGPT-4o 评价在五个维度上的差异，检验结果表 2、表 3 所示。在第一轮中，小组评价与 ChatGPT-4o 评价在五个维度上均未出现显著差异。该现象可能源于我们为 ChatGPT-4o 与小组均提供了明确且清晰的评分标准。在评估过程中，清晰的评分标准发挥了关键的导向作用，促使二者均依据既定规则开展评价，从而使得评价质量未出现显著差异。已有研究也表明，量规是反馈形成的一个关键特征。本轮中小组评价得分略优于 ChatGPT-4o 评价的现象可能与小组成员之间的协作讨论密切相关，集体智慧的碰撞丰富了评价内容，提升了评价的质量，从而使得小组提供的评价结果略优于 ChatGPT-4o 评价。第二轮中，小组评价与 ChatGPT-4o 评价在建设性（$p<0.001$）、准确性（$p=0.010$）和具体性（$p=0.021$）三个维度上出现显著差异，表现为 ChatGPT-4o 评价的得分更高，评价内容更为详尽、可操作，例如：“方案具有较高的实用性，设计了详细的场馆布局和功能分区，满足了多样化的教学需求和环境。特别是智能导览和互动体验区设计，能够显著提升教学效率和学习体验。然而，某些高科技设备的实际操作和维护可能存在一定复杂性，需要进一步验证其可行性”。该结果表明，在第二轮实验中，ChatGPT-4o 的评价质量超越了小组评价，这与 ChatGPT-4o 所具备的迭代特性及自我优化机制密切相关，经由首轮评价活动的学习，ChatGPT-4o 实现了评价质量的大幅提升。尽管协作评价形式对于高质量评价内容的形成具有积极作用，但学生小组的学习能力弱于 ChatGPT，这使得本轮中小组评价质量低于 ChatGPT-4o 的评价质量。这一现象暗示，在教育实践领域，教师可将二者有机融合，以加速小组学习的进程，进而提升教育活动的效率与质量。

表 2 第一轮实验的 Mann-Whitney U 检验结果

质量维度	评价来源	数量	平均秩	秩的总和	Mann-Whitney U	Z	p

有效性	小组	39	40.82	1592.00	709.000	-0.578	0.563
	ChatGPT-4o	39	38.18	1489.00			
建设性	小组	39	40.78	1590.50	710.500	-0.544	0.586
	ChatGPT-4o	39	38.22	1490.50			
准确性	小组	39	43.23	1686.00	615.000	-1.624	0.104
	ChatGPT-4o	39	35.77	1395.00			
完整性	小组	39	41.17	1605.50	695.500	-0.717	0.473
	ChatGPT-4o	39	37.83	1475.50			
具体性	小组	39	36.42	1420.50	640.500	-1.294	0.196
	ChatGPT-4o	39	42.58	1660.50			

表 3 第二轮实验的 Mann-Whitney U 检验结果

质量维度	评价来源	数量	平均秩	秩的总和	Mann-Whitney U	Z	p
有效性	小组	39	35.44	1382.00	602.000	-1.796	0.073
	ChatGPT-4o	39	43.56	1699.00			
建设性	小组	39	32.17	1254.00	474.500	-3.304	< 0.001
	ChatGPT-4o	39	46.83	1826.00			
准确性	小组	39	33.78	1317.50	537.500	-2.567	0.010
	ChatGPT-4o	39	45.22	1763.50			
完整性	小组	39	37.76	1472.50	692.500	-0.783	0.433
	ChatGPT-4o	39	41.24	1608.50			
具体性	小组	39	34.15	1332.00	552.000	-2.315	0.021
	ChatGPT-4o	39	44.85	1749.00			

4.2. 学生对小组评价和 ChatGPT-4o 评价的看法

访谈数据的主题分析结果详见表 4 和表 5。学生们普遍认为小组评价具备高度的完整性和针对性，具体表现在评价内容涵盖了设计方案的各个方面，如创意、表达、逻辑和格式等。此外，评价还针对方案的优点和缺点提供了清晰的反馈意见。小组评价具备此优势的原因与小组之间的协作知识建构以及学生的专业知识背景密切相关。协作过程促使他们汇聚多元化的评估要点，全面覆盖方案内容的诸多方面，从而构建起一个全方位、多视角的综合评价体系。这与 Yukawa(2006)的发现一致，他指出小组成员通过共享知识，共同理解和分析不同观点，从而做出更清晰、准确的决策。然而，学生们也指出，小组评价存在一定的主观性。这可能源于两个因素：首先，个人的情绪和经历可能会影响评估过程(Panadero, 2016)；其次，评估分数对最终成绩的影响可能带来潜在偏见，进而影响客观性。

相较而言，学生普遍认为 ChatGPT-4o 的评价更具建设性和客观性，为方案的修改提供了明确可操作的修改建议，且评价内容较为客观，准确的反映了方案的真实水平。建设性优势可能源于 ChatGPT 丰富的跨学科知识，这使得它能够在处理开放式任务时整合来自不同学科的见解，从而提供清晰、可操作和具体的建议。以往的研究也发现，ChatGPT 可以提供具体而详细的反馈，包括错误识别和问题解决(Ouyang et al., 2024)。客观性则源于 Chat GPT 的机器属性，使得其在评价时不受到外界因素的干扰。与小组评价相同的是，Chat GPT-4o 的评价也具备针对性的优势，这与其敏锐的洞察力有关。尽管 ChatGPT 评价具备上述优势，但其也不可避免的存在机械化的弊端。这主要是源于 AI 系统的固有性质，它依赖于固定的表达模式和评估框架，体现出模板驱动的弊端。

表 4 学生对小组评价的看法

主题	定义	举例

完整性	反馈涵盖工作的所有方面，包括创意、表达、逻辑和格式。	反馈 1 很全面，提到了我们方案的技术特点、实用性、优势、成本、用户需求和格式（第 6 组学生 1）。
针对性	反馈清晰、明确、有针对性，准确反映了工作的真实水平。	反馈 1 准确地指出了我们方案中的优缺点和特色之处，包括一些细节问题（第 7 组学生 2）。
主观性	反馈受个人情绪、经验和偏见的影响。	反馈 1 可能受到个人情感因素或最终成绩的影响，一些评论和分数不那么客观和准确。（第 7 组学生 3）

表 5 学生对 ChatGPT-4o 评价的看法

主题	定义	举例
建设性	反馈根据工作的优势和劣势提供明确、具体可操作的改进建议和想法。	反馈 2 包括详细的改进建议，例如提供全面的用户手册（第 1 组学生 2）。
针对性	反馈清晰、明确、有针对性，准确反映了工作的真实水平。	反馈 2 准确捕捉了方案的亮点和不足，并提供了详细的改进建议（第 4 组学生 3）。
客观性	反馈根据评分标准准确反映了工作的质量和水平。	反馈 2 严格按照评分标准进行评分，我认为它非常客观（第 8 组学生 3）。
机械性	反馈遵循固定的格式和措辞，缺乏特异性和个性化。	反馈 2 中的措辞似乎有些僵化，就像是由机器生成的，缺乏灵活性（第 5 组学生 1）。

5. 结论

本研究全面比较了小组评价与 ChatGPT-4o 评价的质量。结果显示，在第一轮实验中，两种评价方法的质量水平相当。然而，第二轮实验中，ChatGPT-4o 评价在建设性、准确性和具体性方面表现出显著优势，突显了其迭代学习与自我改进的能力。通过访谈数据的主题分析，我们发现学生普遍认为小组评价全面且具有针对性，但同时存在主观性过强的局限性。相比之下，ChatGPT-4o 的评价因其建设性、客观性和针对性受到肯定，尽管在部分场景中可能显得模板化。这些发现为教学实践与评价分析提供了宝贵的洞见。

总体而言，ChatGPT-4o 在高等教育评价领域展现出巨大的潜力。一方面，它可以作为小组评价的有效补充，在小组评价面临困难或主观性限制时，提供更客观、专业且富有建设性的参考意见，帮助小组突破思维瓶颈并提升评价质量。另一方面，教师可以利用 ChatGPT-4o 的迭代学习能力，引导学生学习其分析问题与提供反馈的方法，从而加速小组学习进程，优化整体评价活动效果，推动高等教育评价向多元化、智能化方向发展。未来研究应进一步探索 ChatGPT-4o 与小组评价结合的长期效果，并考察其在不同教育情境中的适应性，以挖掘更广泛的教育价值。

参考文献

孙明霞,岳丹丹.在教学过程中引入同伴评价[J].教学与管理,2012,(25):27-28.

李毅,郑鹏宇,张婷.ChatGPT 赋能教育评价变革的现实前提、作用机理及实践路径[J].现代远距离教育,2024,(03):9-17.

苗逢春.生成式人工智能技术原理及其教育适用性考证[J].现代教育技术，2023(11):5-18.
Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: peer-generated or AI-generated feedback? International Journal of Educational Technology in Higher Education, 21(1), 23.
Barrot, J. S. (2023). ChatGPT as a language learning tool: An emerging technology report. Technology, Knowledge and Learning, 1–6. https://doi.org/10.1007/s10758-023-09711-4
Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.
Chen, W., Tan, J. S., & Pi, Z. (2021). The spiral model of collaborative knowledge improvement: An exploratory study of a networked collaborative classroom. International Journal of Computer-Supported Collaborative Learning, 16, 7–35.
Ching, Y. H., & Hsu, Y. C. (2016). Learners' interpersonal beliefs and generated feedback in an online role-playing peer-feedback activity: An exploratory study. International Review of Research in Open & Distance Learning, 17(2), 105–122. https://doi.org/10.19173/irrodl.v17i2.2221
Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. Education and Information Technologies, 29(1), 8435–8463.
Guo, K., Chen, X., & Qiao, S. (2022). Exploring a collaborative approach to peer feedback in EFL writing: How do students participate? RELC Journal, 00336882221143192.
Hsia, L. H., Huang, I., & Hwang, G. J. (2016). Effects of different online peer-feedback approaches on students' performance skills, motivation, and self-efficacy in a dance course. Computers & Education, 96, 55–71. https://doi.org/10.1016/j.compedu.2016.02.004
Joordens, S., Pare, D. E., & Pruesse, K. (2009). PeerScholar: An evidence-based online peer assessment tool supporting critical thinking and clear communication. In Proceedings of the 2009 International Conference on e-Learning (pp. 236-240).
Katz, A., Wei, S., Nanda, G., Brinton, C., & Ohland, M. (2023). Exploring the efficacy of ChatGPT in analyzing Student Teamwork Feedback with an existing taxonomy. arxiv preprint arxiv:2305.11882. https://doi.org/10.48550/arXiv.2305.11882.
Kleinman, Z. (2024, May 14). OpenAI's new model GPT-4o can teach maths and flirts - but still glitches. BBC. https://www.bbc.com/news/articles/cv2xx1xe2evo
Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? Assessment & Evaluation in Higher Education, 49(5), 616–633. https://doi.org/10.1080/02602938.2024.2301722
Noroozi, O., Biemans, H., & Mulder, M. (2016). Relations between scripted online peer feedback processes and quality of written argumentative essay. The Internet and Higher Education, 31, 20–31. https://doi.org/10.1016/j.iheduc.2016.05.002
Ouyang, F., Guo, M., Zhang, N., Bai, X., & Jiao, P. (2024). Comparing the effects of instructor manual feedback and ChatGPT intelligent feedback on collaborative programming in China's higher education. IEEE Transactions on Learning Technologies.
Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: A review and future directions. In G. T. L. Brown, & L. R. Harris (Eds.), Handbook of human and social conditions in assessment (pp. 247–266). New York: Routledge.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., ... & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. Learning and Instruction, 91, 101894.

Tan, J. S., & Chen, W. (2022). Peer feedback to support collaborative knowledge improvement: What kind of feedback feed-forward? Computers & Education, 187, 104467.

Topping, K. (1998). Peer assessment between students in colleges and universities. Review of Educational Research, 68(3), 249–276. https://doi.org/10.2307/1170598

Topping, K. (2009). Peer assessment. Theory into Practice, 48(1), 20–27

Topping, K. (2017). Peer assessment: Learning by judging and discussing the work of other learners. Interdisciplinary Education and Psychology, 1(1), 1–17.

Villamil, O. S., & Guerrero, M. C. D. (1998). Assessing the impact of peer revision on L2 writing. Applied linguistics, 19(4), 491-514.

Wang, X. M., Hwang, G. J., Liang, Z. Y., & Wang, H. Y. (2017). Enhancing students' computer programming performances, critical thinking awareness and attitudes towards programming: An online peer assessment attempt. Journal of Educational Technology & Society, 20(4), 58–68

Wiggers, K. (2024). OpenAI debuts GPT-4o'omni'model now powering ChatGPT. TechCrunch. Retrieved May, 16, 2024.

Yukawa, J. (2006). Co-reflection in online learning: Collaborative critical thinking as narrative. International Journal of Computer-Supported Collaborative Learning, 1, 203–228.

Zhang, S., Li, H., Wen, Y., Zhang, Y., Guo, T., & He, X. (2023). Exploration of a group assessment model to foster student teachers' critical thinking. Thinking Skills and Creativity, 47, 101239.

Zheng, L., Chen, N-S., Li, X., & Huang, R. (2016). The impact of a two-round, mobile peer assessment on learning achievements, critical thinking skills, and meta-cognitive awareness. International Journal of Mobile Learning and Organisation, 10(4), 292-306. https://doi.org/10.1504/IJMLO.2016.079503