人工智能能否超越优等生解决高中化学问题？：基于能力测评的证据

Can Artificial Intelligence Surpass Top Students in Solving High School Chemistry Problems? Evidence from Competency Assessment

陈玲芳[1]，薛松[1*]
[1]浙江师范大学教育学院
[*]xuesong@zjnu.edu.cn

【摘要】 生成式人工智能的快速发展为教育带来了新的机遇与挑战。为了评估生成式人工智能在高中化学教育中的应用潜力，研究选取部分化学试题对 ChatGPT、讯飞星火两大生成式人工智能的解题能力进行了测试，并与重点中学选考化学的高三学生的解题能力进行了比较。结果显示两大模型准确率仅高于小部分学生。进一步分析错误类型后发现，生成式人工智能在解题过程中存在概念理解错误、信息提取不足等问题。基于此，研究为生成式人工智能辅助化学教育及其自身未来的优化与发展提供了启示与建议。
【关键词】 生成式人工智能；ChatGPT；讯飞星火；高中化学；问答测试

Abstract : *The rapid development of artificial intelligence generated content (AIGC) presents new opportunities and challenges for education. this study tested the problem - solving abilities of ChatGPT and iFlytek Spark on chemistry questions, comparing them with those of senior high - school students taking chemistry in key schools. Results show that the accuracy rates of the two models are only higher than those of a small number of students, with students' performance significantly better. Further analysis of error types reveals problems like conceptual misunderstandings and insufficient information extraction in the models' problem - solving. Based on this, the study provides insights and suggestions for the use of generative AI in chemistry education and its future optimization.*
Keywords: AIGC, ChatGPT, iFLYTEK Spark, Chemistry in High School, Quiz Test

1.前言

推出生成式人工智能产品 ChatGPT 以来，以 ChatGPT 为代表的聊天机器人在教育领域的应用迅速成为关注焦点。由于生成式人工智能可以使用自然语言处理技术生成逻辑合理的回答，许多学者研究了其解答化学问题的潜力。然而，研究主要聚焦于大学化学领域（Clark et al.,2023；Clark et al.,2023；Watts et al.,2023），在中学化学的应用能力未得到充分评估，国内相关研究缺乏，仅有学者对生成式人工智能解决物理（刘丹等，2024）、生物（闫白洋等，2024）等其他科学问题能力进行研究。因此，针对生成式人工智能在中学化学教育中的应用展开探讨，不仅能够填补现有研究空白，还能为探索其在国内高中化学教学中的潜在价值提供新视角。研究选取不同版本的 ChatGPT 和讯飞星火，比较了它们与重点中学学生在解答高中化学问题时的准确率，分析了生成式人工智能在解题过程中影响正确率的错误类型，旨在评估生成式人工智能在高中化学教育中的应用潜力。

2.研究设计

2.1. 测试对象

测试于 2024 年 6 月-8 月进行，选择 ChatGPT 和讯飞星火作为测试平台。ChatGPT 有 ChatGPT3.5、ChatGPT4.0、ChatGPT4o 三个版本被广泛使用，可通过访问“ChatGPT (openai.com)”进入 ChatGPT 测试平台。用户可以通过“讯飞星火认知大模型-AI 大语言模型-星火大模型-科大讯飞 (xfyun.cn)” 免费使用讯飞星火平台。为了比较 ChatGPT、讯飞星火两大生成式人工智能与学生在解决高中化学问题的表现，研究选取浙江省某重点中学的高三学生为测试对象。参与本次测试的学生分为三类：第 I 类为成绩优异的学生，共 98 人，这些学生具备较强的学科理解和问题解决能力，且多数参与化学竞赛；第 II 类为选考物化组合科目的学生，有扎实的理科基础，共 113 人；第 III 类为选考化学与文科（历史、政治或地理）科目的学生，共 29 人。统计选考化学并实际参与测试的学生人数，共计 240 名有效样本。

2.2. 测试内容与方法

浙江省名校协作体试题是由多所学校联合命题、多名教育专家共同参与所编制的一套试卷。该试题为高三年级开学考试化学试题，涉及常见无机物及其应用、有机化学基础等内容。该卷难度适中，信效度良好，且数据较易获取。尽管生成式人工智能在图片识别方面具备一定能力，但在分析化学相关图片时仍存在识别不准确等局限。而化学测试题中的某些图片信息无法用文字精准描述，为尽可能排除生成式人工智能识别图片信息对解题产生的干扰因素，本研究所选取的题目为试题中不包含图片的选择题部分。题目以文字叙述为主，符合标准共计 17 道。

研究采用指令式提问的方式进行测试，提问指令为“请以你现有化学知识分析并计算，逐个分析选项判断其正误，选择出唯一一个最准确答案，并指出其他选项错误的地方”。如果生成式人工智能仍给出两个或两个以上正确答案，则进一步对每个选项逐个提问，让其根据现有化学知识分析、计算、判断其正误。若生成式人工智能仍然认为有两个或两个以上答案，则判定其做错了该题。本次测试所得数据均采用此方法得出，未使用其他方法进行提问。

3.研究结果与分析

3.1. 答题结果与分析

对样本中的 17 道不含图片的题目进行测试后，结果如表 1 所示 ，ChatGPT3.5 正确回答了 8 题，正确率为 47.06%。相比之下，ChatGPT4o 答对了 9 题，显示出新一代工具在解决高中化学无图客观题方面略有提升。对比分析三代 ChatGPT 的回答，发现 ChatGPT4o 在题意理解和知识调用能力上比上一代工具更强。同时，国内研发的讯飞星火正确解答了 12 题，错误 5 题，正确率达到 70.59%，明显高于 ChatGPT3.5 和 ChatGPT4.0，表现出较强的解题能力。分析学生的答题情况，发现三类学生均展现出了较高的正确率，结果如表 1 所示。特别是I类化学竞赛生，平均正确率高达 88.12%，同时尽管III类学生来自化学和其他学科的混合背景，其解题表现也相当出色，平均正确率达到 81.14%，与I类和II类学生的水平相差不大。

表 1 各模型与各类学生作答情况

	最高正确题数	最低正确题数	正确率
讯飞星火	12	——	70.59%
ChatGPT4o	9	——	52.94%
ChatGPT4.0	9	——	52.94%
ChatGPT3.5	8	——	47.06%
I类	17	10	88.12%
II类	17	9	85.68%
III类	17	10	81.14%

统计答对不同题数的学生人数所占总人数的百分比，并与生成式人工智能的答对情况作比较，如图 1 所示。对于参与测试的重点中学学生来说，有近 10%的学生能正确回答出全部题目，约一半学生作答正确题数集中在 15 题和 16 题，整体展现出了较高的水平，而生成式人工智能与学生之间的表现存在明显差异。尽管讯飞星火在所测评的生成式人工智能中的表现最佳，准确率达到 70%，但与学生相比，大部分学生的解题水平都明显高于生成式人工智能，有高达 90%的学生作答正确率超过 70%。这一结果表明，生成式人工智能的解题能力尚未达到重点高中大多数学生的水平，仍有较大的提升空间。

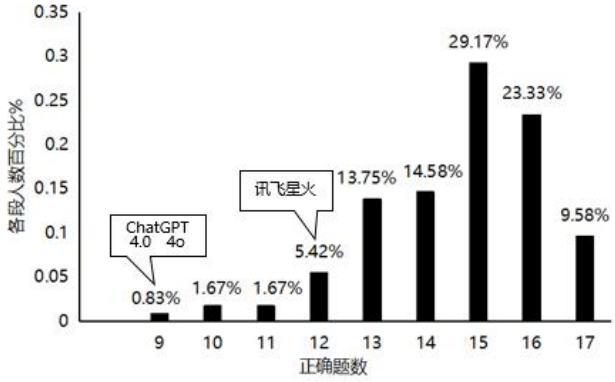

图 1 学生与各生成式人工智能作答情况比较

3.2. 正确回答的解释能力分析

研究将测试题目进一步划分为简单问题和复杂问题，分析比较两款生成式人工智能在正确回答两类问题基础上的解释能力。其中简单问题是指涉及基本概念、基础知识或单一步骤解决的化学问题。通过直接应用化学知识、公式或简单的逻辑推理就可以解决，不需要复杂的分析或高级思维技能。复杂问题是指涉及多个变量和条件，需要综合运用化学知识和理论进行解决的化学问题。这些问题往往不只是停留在简单的知识记忆层面，而是要求学生能够进行深入分析和思考。通过测试，两大模型在解决部分简单问题的优势明显，不仅能够产生正确、完整且流畅的表达，还能对一些基础概念进行深入拓展，帮助学生了解课本以外的知识，完善知识结构。在处理一些复杂问题时，两大模型都显示出较强的计算能力和分析能力。但值得注意的是，对于某些特定的复杂问题，完全可以利用已有的知识进行有效的定性判断，过度依赖定量计算反而可能导致解题过程复杂化。在应对两类问题上，生成式人工智能确实展现出了其特有的优势，但整体作答正确率并未达到预期，在问题解决过程中仍然呈现出多方面的挑战。这些局限性在一定程度上阻碍了大型语言模型在高中化学教学领域的广泛应用。因此，本研究针对讯飞星火和 ChatGPT 的错误回答进行梳理和分析，旨在促进生成式人工智能在高中化学教育领域发挥更大的作用。

3.3. 错误回答的原因分析

通过对生成式人工智能错误回答的深入梳理与分析，发现其作答表现受多个因素的影响，主要集中在以下几个方面：

对化学基本概念和原理的理解不够深入，导致无法准确应用相关知识进行解答。这一点在 ChatGPT 3.5 模型中表现尤为明显。例如， ChatGPT3.5 在探讨草酸（$H_2C_2O_4$）在化学反应中的碳元素价态变化时，虽然 ChatGPT3.5 能够正确识别出碳元素在反应前后的化合价变化，从+3 氧化态变为+4 氧化态（$C_2O_4^{2-} \rightarrow CO_2$），但错误地判断了碳元素是被还原，从而得出草酸是还原剂，产生了“还原剂被还原”的迷思概念。这一错误揭示了 ChatGPT3.5 未能正确理解化合价变化与氧化剂或还原剂之间的关系，导致其在后续分析中对氧化还原反应中氧化剂、还原剂及其产物之间关系的判断完全错误。

未能有效处理化学问题中多因素作用，导致分析结果不准确。例如题目为：“常温下 KNO_3 溶液和 CH_3COONH_4 溶液 pH 均为 7，两溶液中水的电离程度是否相同？” 正确解答此题需要综合分析盐类水解对水的电离平衡造成的影响。尽管 KNO_3和 CH_3COONH_4溶液均表现出 pH 为 7，但二者在水的电离程度上存在明显差异。具体而言，CH_3COO^-和 NH_4^+由于发生水解促进了水的电离，因此 CH_3COONH_4溶液中水的电离程度比 KNO_3溶液更大。然而，ChatGPT 在分析时仅基于溶液 pH 值为 7 这一单一信息，错误地判定水电离出来的 H^+和 OH^-的浓度均为 1.0×10^{-7} mol/L，忽略了 NH_4^+和 CH_3COO^-的水解作用对水电离平衡的实际影响。上述结果反映了 ChatGPT 在处理涉及多因素的化学问题时，倾向于依赖单一信息（或知识点），未能综合考虑其他相关因素，导致对问题的理解不够全面，进而得出不准确的结论。

未能全面提取题目中的关键信息，导致分析结果出现偏差。例如，题目询问“二氧化氮是否能使湿润的淀粉碘化钾试纸变蓝”。ChatGPT 在分析过程中仅关注到 NO_2 本身并非强氧化剂，因而不能将 I^-氧化成 I_2，进而无法与淀粉结合形成蓝色复合物，使试纸变蓝。然而，该模型忽略了题目中的“湿润”这一关键条件。实际上，NO_2 易溶于水，在湿润条件下，可以与水反应生成具有强氧化性的 HNO_3，从而使淀粉碘化钾试纸变蓝。类似的错误在讯飞星火模型中也有发生。

4.研究结论

研究对 ChatGPT3.5、ChatGPT4.0、ChatGPT4o 以及讯飞星火模型解答浙江省名校协作体试题中无图客观题时的正确率进行了测试，发现讯飞星火模型正确率最高，而 ChatGPT4.0 与 ChatGPT4o 的正确率相同，略高于 ChatGPT3.5。这表明生成式人工智能在解答高中化学问题方面具备一定能力，其中讯飞星火的表现尤为突出。然而，将这些模型与重点中学选考化学的高三学生的表现比较时发现，两大模型的解答能力仅优于少部分学生。随着 ChatGPT 版本的更新，其正确率有所提升，这表明 ChatGPT 的性能在不断增强，显示出生成式人工智能在解题方面的巨大潜力。与重点中学学生的表现相比，生成式人工智能的解题能力仍然存在明显不足。通过对错题类型进行分析，发现生成式人工智能在解题过程中存在诸多问题。这些发现为生成式人工智能在辅助化学教学中的应用以及未来优化与发展提供了重要启示。

参考文献

刘丹 & 童大振. (2024). ChatGPT4 解决科学问题能力的探究——以“北上广”三市中考物理试题为例. 中学物理 (08), 15-19.

闫白洋 & 佘建云. (2023). 生成式人工智能在高中生物学教学领域中的问答测试与使用建议. 生物学教学 (09), 34-36.

Clark, T. M., Anderson, E., Dickson-Karn, N. M., Soltanirad, C., & Tafini, N. (2023). Comparing the Performance of College Chemistry Students with ChatGPT for Calculations Involving Acids and Bases. Journal of Chemical Education, 100(10), 3934-3944.

Clark, T. M. (2023). Investigating the Use of an Artificial Intelligence Chatbot with Ggeneral Chemistry Exam Questions. Journal of Chemical Education, 100(5), 1905-1916.

Watts, F. M., Dood, A. J., Shultz, G. V., & Rodriguez, J. M. G. (2023). Comparing Student and Generative Artificial Intelligence Chatbot Responses to Organic Chemistry Writing-to-learn Assignments. Journal of Chemical Education, 100(10), 3806-3817.