

利用大语言模型多智能体探索复杂推理数学题的苏格拉底式教学方法

Socratic Teaching Method for Exploring Complex Reasoning Mathematical Problems Using Large Language Models(LLMs) Multi-agents

张誉月¹, 贾积有^{1*}

¹ 北京大学教育学院教育技术系

persistbetter@stu.pku.edu.cn, *jjy@pku.edu.cn

【摘要】本研究针对数学教学中直接给予答案、缺乏深度引导的问题,提出一种基于大型语言模型多智能体的苏格拉底式教学框架。通过设计7类学生真实学习时的认知状态,整合学生与“教学主管-教师”的多智能体协作,构建了动态多轮对话系统与引导性提问,模拟真实教学场景中逐步提问、纠错与反思的过程。基于30道高等数学题的实验表明,模型在错误答案识别(84.73)与成功拒绝率(85.10)上表现突出,但成功解释率(82.47)因逻辑复杂性受限。研究证实,该方法能有效引导学习者主动思考,通过错误根源挖掘提升问题解决能力。未来的研究将进一步扩展数据集规模,优化智能体反馈机制等。

【关键词】智能教学系统;高等数学教育;大语言模型;苏格拉底式教学;多智能体系统

Abstract: This study proposes a Socratic teaching framework based on a large language model multi-agent system to address the problem of direct answers and lack of deep guidance in mathematics teaching. By designing the cognitive states of seven types of students during real learning, integrating the multi-agent collaboration between students and "teaching supervisors teachers", a dynamic multi round dialogue system and guided questioning were constructed to simulate the process of gradually questioning, correcting errors, and reflecting in real teaching scenarios. An experiment based on 30 advanced mathematics problems shows that the model performs outstandingly in identifying incorrect answers (84.73) and achieving a successful rejection rate (85.10), but the successful interpretation rate (82.47) is limited by logical complexity. Research has shown that this method can effectively guide learners to think proactively and improve their problem-solving abilities by identifying the root causes of errors. Future research will further expand the dataset size and optimize the feedback mechanism of intelligent agents.

Keywords: Intelligent Tutoring System(ITS); Higher mathematics education; Large Language Models (LLMs); Socratic teaching; Multi-agent system

1.引言:生成式人工智能在数学解题中的局限

大型语言模型(LLM)在自然语言处理、翻译和问答等各种任务中取得了令人瞩目的成果(Wang等,2023)。这引起了人们的广泛关注,即使用LLM来彻底改变智能教育,尤其是个性化教学的潜力(Ma等,2024;Adıgüzel等,2024),这主要归功于它们的两个优势。一方面,LLM已经表现出对数学和物理等基础学科知识的类似人类的掌握和熟练程度

(Achiam等,2023)。因此,它们可以像真正的老师一样有效地向学生提供特定学科的指导。另一方面,LLM对用户(即学生)的输入表现出卓越的理解和适应性。它们能够理解不同学生的需求,提供即时反馈并参与互动教学。

基于这两个方面,已经有很多研究将通用的LLM(例如ChatGPT)应用于个性化教学(Chen等,2023),或构建特定的教学LLM,例如MathGPT2、EduGPT3和EduChat(Dan等,2023)。然而,目前基于LLM的个性化教学方法主要遵循“问答”范式,被动地提供诸如提供问题答案和向学生解释知识概念等功能。在这个过程中,他们将教学过度简化为一系列问答,直接根据CoT,ToT等提供完整的答案(Wei等,2022;Yao等,2024),这不足以真正识别学生可能存在的问题并提供有针对性的帮助。因此,学生可能难以理解解决问题的过程,能力得不到真正的提高,并且无法在未来解决类似的问题。

数学是一种需要通过结构化符号和系统进行复杂推理的语言,与口语规则相似,是人类智能的一项重要能力。近年来,利用人工智能技术自主解决数学问题已引起人们的关注,早在1963年就已开始(Briars和Larkin,1984)。关于数学的人工智能研究分为三部分:基于统计学习的方法(Zhou等,2015)、传统机器学习技术(Kushman等,2014)和基于深度学习的方法(Wang等,2017)。近年来,大型语言模型(LLM)在数学领域取得了巨大成功,

出现了各种类型的数学数据集 (Cobbe 等, 2021; Zhao 等, 2020) 和数学 LLM (Xu 等, 2023; Yu 等, 2023; Yue 等, 2023)。然而, 以往对数学 LLM 的研究主要侧重于提高数学能力, 其在教学中的应用仍然有限。LLM 在面对数学问题时倾向于直接给出解决过程, 或分析总结提炼文本。Jia 等人基于已有数据集分别让人类教师 and 大型语言模型生成若干组通用型提示, 将提示混合后让专家打分, 统计分析结果表明人类专家无法区分人工和机器产生的通用性提示 (Jia 等, 2024)。关于数学辅导的应用主要集中于 K12 教育, 对于高等数学等更加复杂的数学题很少有人研究。

在本文中, 我们借鉴了苏格拉底式教学法 (Paul 和 Elder, 2007), 提出了一种基于大型语言模型多智能体的苏格拉底式教学框架。它实现了一种新颖的教学范式, 该范式的关键是让学生参与对话, 积极参与学习过程, 不断提出开放式问题, 鼓励他们表达自己的想法, 挑战假设, 独立思考。

2. 相关研究

2.1. 苏格拉底教学的定义和优势

苏格拉底教学的关键是让学生参与对话, 积极参与学习过程, 鼓励他们表达自己的想法, 挑战假设, 独立思考。这个过程使学生学会自己解决问题, 从而培养更深层次的掌握和能力。苏格拉底式提问允许教师定制他们自己的问题, 以适当地挑战学生, 同时也与学生个人的解决问题的能力保持一致 (Al-Hossami 等, 2023)。苏格拉底教学的理念是知识不是简单地传递, 而是通过动态的探究和对话过程发现的。有两个基本要求: 1) 它本质上是对话式的, 依靠教师和学生之间的对话来促进学习; 2) 它使用探索性问题来积极吸引学生, 促进独立思考并鼓励他们自己寻找答案。

精心设计的苏格拉底式问题可以鼓励学生自我解释。提高他们对任务的理解, 促进有效规划解决方案 (Micheline 等, 1994)。此外, 它们还可以帮助发现学生知识方面的重大差距。这种范围使教育工作者能够设计出具有适当挑战性但又在学生回答能力范围内的问题。

2.2. 优秀的苏格拉底提问策略

尽管苏格拉底教学存在诸多优势, 但需要注意谨慎使用提问策略, 因为不适当的内容有时会弊大于利, 例如若提出总是容易回答的问题不会导致更深入、更广泛的学习。对参与者而言, 对于不需要概念理解的子问题或对于更复杂的问题, 可能会使任务看起来更具挑战性。

优秀的提问策略有两个特征: 首先, 重点突出: 良好提问策略的一个基本属性是针对最关键的特定领域内容提出问题。无关的问题不仅使过程变得困难, 还会迫使注意力转移, 并可能增加学生的认知负担。其次, 目标驱动: 以正确的顺序提出相关问题, 帮助学生达到最终目标 (如果是数学应用题, 则解决主要问题), 这是良好提问的另一个重要部分 (Shridhar 等, 2022)。

此外, 一些比较宽泛的问题, 例如真实教学对话场景中常常会要求学习者解释 (例如: 你能重新表述吗?)、提供例子 (例如: 你能提供一个例子吗?) 和暗示 (例如: 你为什么认为……?) 之类开放性的问题, 尽管学生可能回答地不尽如人意或回答错误, 这类对话也很有意义, 因为学习者的错误和误解可能为教师有效地了解学习者的认知状态提供了一个重要的诱饵, 能够使得教师和学习者共同负责推动对话。

2.3. 大语言模型下的苏格拉底教学法

苏格拉底式的智能辅导系统的适用性, 往往受到针对特定练习的苏格拉底式话语的预定义和手动定制的限制 (Alshaikh 等, 2020)。Al-Hossami (2020) 提出了一个由苏格拉底式对话组成的数据集, 旨在帮助新手程序员纠正基本计算问题中的错误。Shridhar 的研究探索了用于教育目的的数学问题解决方案的自动生成策略。Qi 等人 (2023) 将苏格拉底式提问视为一种递归思维过程, 它将复杂问题分解为更简单、相关的子问题。

然而, 先前的研究发现主要存在以下几个问题: 第一, 无法保证答案的可靠性和质量, 在幻觉问题的复杂推理上表现不佳。第二, 教学策略方面, 用怎样的问题来指导和启发学生的策略 (何时提问、如何提问和问什么) 尚不明确, 太多或太少太难或太容易的提示都会影

响学生的学习过程。第三, 先前与数学教学有关的智能辅导系统研究主要集中在 K12 领域, 缺乏高等数学等复杂推理数学题的教学应用。而对于复杂推理数学题存在更多的难点, 如对 LLM 解题质量的要求更高、复杂数学公式与 latex 格式的转化等。

为此, 我们使用多 agent 的优秀提问策略来生成高等数学教育领域的个性化苏格拉底提示, 以解决幻觉问题、教学策略问题等复杂推理数学题下诸多难点。

3. 基于大型语言模型多智能体的苏格拉底式教学设计

3.1. 多智能体综合方案

为解决复杂推理数学题中题目更困难、幻觉现象可能更严重等问题, 我们使用一种新颖的“教学主管-教师-学生”综合解决方案来生成优势的苏格拉底式师生对话, 该综合解决方案由三个 LLM 智能体组成。每轮教学对话都是在教学主观的监督下, 由教师和学生交互循环形成的, 每个 agent 都用 GPT4o 进行模拟。

1) 教师智能体

教师智能体积极地以苏格拉底式风格激励学生解决问题, 首先, 它会在适当的时候用苏格拉底式问题提示学生思考, 比如在完成一个推理步骤后引导学生考虑下一步。其次, 它需要向学生提供步骤和所涉及的知识点的解释。最终, 如果教师智能体认为教学过程已完成, 它将输出 “[END]” token 作为其输出的结尾, 表示终止循环。

2) 学习者认知状态模拟

在教师提问的过程中, 我们模拟真实学生对教师智能体的提示 (即问题和解释) 对教师智能体进行回复。为了确保学生回复的真实性和多样性, 我们从学生的角度回顾了苏格拉底式教学过程: 首先, 学生需要掌握手头问题的含义; 然后理解老师提供的指示, 并利用计算能力和获得的知识来执行指令; 最终激发学习兴趣, 并有助于培养有效的学习成果。基于这一思想, 并综合文献综述的结果, 我们总结了在实际教学过程中, 7 种真实学生的回复的分类, 并在每次回复时模拟其中的一种。学生的反应和认知状态可以分为以下 7 类:

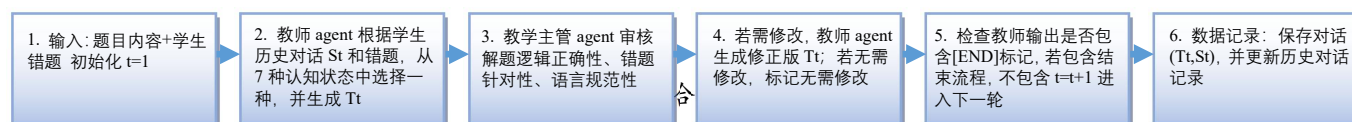
表 1 学生可能的 7 类反应状态

序号	状态类型	含义
1	问题理解错误	没有清楚地理解问题或给出的说明
2	指令理解错误	学生难以理解和执行教师指令的程度。处于良好状态的学生应该很容易完成这些指令。
3	计算错误	学生错误推导数学表达式和数字的能力
4	知识未掌握	学生未掌握定理、概念等理论知识。
5	学习动机不足	学生寻求和获取新信息、提出问题和探索可能性的愿望或倾向很低。
6	不相关	学生的输入与教学内容无关, 例如问“今天天气怎么样?”
7	正确回答	学生正确回答了老师的问题, 可以继续推进。
8	部分回答	提供了预期答案的一部分

此外, 为了获得关于学生错误类别的最佳提示, 对于学生的错误回答, 我们定义了几个具体的错误分类, 包括: 积分时忘记常数项的处理、积分范围不正确、混淆无穷级数的收敛条件、误用洛必达法则、误解泰勒展开的应用、代数运算中的符号错误等。

3) 教学主管智能体

由于 GPT 在理解学生和担任教师所需的语言表达方面依然存在不足。为了解决这个问题, 我们提出了一个教学主管 agent 充当监督角色, 判断教师的指示是否符合苏格拉底式教学的要求。如果它认为指示不符合要求, 它有权在指示提交给学生之前对其进行修改。重点关注: 1、解题思路是否正确; 2、是否符合苏格拉底式风格; 3、是否针对学生的错误进行针对性解答; 4、其语言风格是否类似于真正的老师, 是否简短明确。若符合要求, 则返回“无需修改”。



以图 1 为例，在第一轮 ($t=1$) 中，教师 agent 直接根据输入的题目和学生的错题情况以苏格拉底式的方式（不提供答案，而是遵循逐步提问）提出第一个问题 T1。在教师 agent 生成 T1 后，教学主管 agent 对其进行判断并修改，如果判断需要修改，则修改后的响应随后发送给教师 agent，若判断无需修改，则返回“无需修改”。

然后，我们从 7 种认知反馈中选择一个状态，并模仿学生与教师 agent 进行对话。我们的回答作为第二轮 ($t=2$) 的输入给教师 agent，教师 agent 继续根据此前的对话和新的输入，以苏格拉底式提供指导 T2。修改后的响应随后发送给 Student，开始下一轮对话。最终，如果 Teacher 认为教学过程已完成，它将输出 “[END]” token 作为其输出的结尾，表示终止循环。这样，使得每轮教学对话 $\{(T1, S1), (T2, S2), \dots\}$ 中的内容，都是在教学主管的监督下，由教师和学生交互循环形成的。

3.2 提示编写方法

提示词质量对于生成通用型提示的效果有着显著的影响。作为生成通用型提示的关键变量，不同类型的提示词对模型的理解和输出结果产生不同的效果。大模型提示词的构造方法多样，包括零样本学习 (Zero-Shot Learning)、单样本学习 (One-Shot Learning) 和少样本学习 (Few-Shot Learning) 等。Jia 等人的研究表明单样本和少样本学习会导致 GPT 过度依赖给出的例子，而忽略了数学问题的具体特点和变化，而零样本学习使 GPT 能够根据具体问题自由生成提示，产生的通用型提示得分最佳 (Jia 等, 2023)。因此本文主要采用零样本学习的方法，提示如下：

表 2 复杂推理数学题的苏格拉底式引导型提示词

教师 agent	请您扮演一名大学高数课教师，您将会收到一道高数题目、该题目正确的解答过程和高数初学者错误的推理过程。目标是：1、通过多轮对话和提示，帮助高数初学者纠正练习中的推理错误，必须采用苏格拉底式提问的形式，引导学生一步步得出正确答案。此外，提示不得包含练习或其中任何部分的正确答案。2、当学生最终得出正确结果时，对高数初学者练习中错误对学生的错误进行分类。题目：【题目】；学生错误的答题过程：【答题过程】。
教学主管 agent	请您扮演一名主管，判断教师的提示是否符合苏格拉底式教学的要求。 1、如果【学生】问的是常识，或者很难理解某个概念，【老师】可以直接解释。2、其他情况下，不能以陈述或修辞的方式给出问题的直接答案或解决方案。3、每次只能提 1 个问题，即最多只能有 1 个“？”存在。 若符合要求，返回无需修改，若不符合要求，指出如何进行修改。注意只针对这一条提示提建议，不需给出学生可能回答后的后续提示。教师的提示为：【教师提示】。
教师 agent (修改)	根据主管的反馈，模仿与优秀的高数教师，修改您的原提示，主管的反馈是：【教学主管反馈】。

4. 实验结果

4.1 数据收集和实验流程

本研究所收集的高数问题主要来自于近 5 年来真实的考研数学题，以及辽宁工学院数理科学系编著的《高等数学学习中的常见错误及评析》，其中包含了 30 道高数题目、代表性错误解法、错误类型以及正确解法。这些数学题涵盖了各个难度等级的题目，而代表性错误往往具有一定的普遍性，包括中国高数初学者在对概念和定理的理解不深，逻辑推理不严密，计算上不顾公式应用条件盲目套用等诸多方面的错误 (辽宁工学院, 1995)，具有权威性和教学指导性。为了更方便快捷地调试 prompt，使用 Python Streamlit 框架搭建了 web 可视化和交互平台，如下图所示：



图 2 web 可视化和交互平台

与 K12 数学题目不同，复杂推理数学题中包含大量复杂的公式，为解决该问题，我们首先将 pdf 文件中的题目分割为图片，接着调用 api 将图片中的公式识别为 latex 格式，接着转化为 markdown 格式，上传给大语言模型。大语言模型返回的 latex 格式再调用 api 转化为可视化的公式。

4.2 定性测评结果

针对 30 道题目，我们依次对每道题目进行了对话，所有题目结束之后，对话自动保存，用于后续的分析。由于篇幅限制，我们仅展示其中一道题目的部分问答内容，其中教师 agent 为该系统给出的反馈，学长则为某大一真实高数学习者：

表 3 某例题的苏格拉底式问答

题目：“求极限： $\lim_{x \rightarrow 0} \frac{x^2 \sin \frac{1}{x}}{\sin x}$ ”	【错误解法】 $\lim_{x \rightarrow 0} x^2 \sin \frac{1}{x} = \lim_{x \rightarrow 0} x^2 \cdot \lim_{x \rightarrow 0} \sin \frac{1}{x} = 0 \cdot \lim_{x \rightarrow 0} \sin \frac{1}{x} = 0$
教师 agent：第一步： 你能告诉我，为什么你认为在 $x \rightarrow 0$ 时该表达式是 0/0 型未定式吗？	
学生：因为 x 趋近于 0 时，分子和分母的极限都为 0。	
教师 agent：很好，你正确地指出了分子和分母在 $x \rightarrow 0$ 时都趋近于 0。但是，在应用洛必达法则之前，分子和分母的导数存在并且连续。你能尝试考虑分子中的 $x^2 \sin \frac{1}{x}$ 在 $x \rightarrow 0$ 时的导数吗？	
学生：似乎 $\sin(1/x)$ 不能求导	
教师 agent：非常好，你注意到了关键点！考虑到这一点，你认为我们是否可以继续使用洛必达法则？	
……（由于篇幅原因省略）	
教师 agent：非常好！你总结得很准确。总结： 洛必达法则的使用条件：必须确保分子和分母在考虑的点具有可求导性。	
夹逼定理的应用：通过分析分子 $x^2 \sin \frac{1}{x}$ 的极限，使用 $\sin \frac{1}{x}$ 在 $[-1,1]$ 范围内的性质，可以推导出分子的极限是 0。 小角度近似：通过将 $\sin x \approx x$ 用于分母，可以进一步简化问题，最终得出正确的极限结果。	

可以看出，本研究利用多智能体生成的苏格拉底式教学法具有以下优点：首先，引导性强，通过问题鼓励学生思考，让学生自己发现问题，深入理解概念。第二，逻辑结构清晰，教师 agent 将问题分为多个步骤，每一步的解释和分析都紧密相连，形成了清晰的思路。最后，提供总结与反馈。教师 agent 会对学生提出的问题进行了回顾和归纳，并帮助学生总结了如何避免类似的错误，这对于学生数学学习的触类旁通非常重要。

4.3 定量测评结果

为了评价本研究生成的苏格拉底教学方法的效果，我们通过以下评价体系对其进行评价，涵盖苏格拉底风格和教学能力五个教学维度：

表 4 苏格拉底对话的评价维度

	评价维度	含义
1	总体质量	对教学质量的整体和主观评估，要求教学满足苏格拉底风格并增强学生的体验。
2	错误答案识别准确率	关注老师是否能准确识别学生的“错误回答”。
3	正确答案识别准确率	与错误识别相反，该维度关注模型是否能够准确识别学生的“正确回答”。
4	成功解释率	是否能为学生的“提问”提供令人满意的解释。

5	成功拒绝率	针对教师拒绝回答学生“不相关”问题并将其重定向回教学内容的情况而设计的。
---	-------	--------------------------------------

我们将 30 道题目的对话依次让 gpt-4o 进行打分，满分均为 100 分，所得到的结果如下：
表 5gpt-4o 对 30 道题目的苏格拉底对话打分结果

	30 题平均值	标准差	中位数	最大值	最小值
总体质量	83.40	7.66	84.50	99.00	70.00
错误答案识别准确率	84.73	8.03	84.00	99.00	70.00
正确答案识别准确率	84.50	9.28	84.00	99.00	70.00
成功解释率	82.47	8.44	82.50	100.00	70.00
成功拒绝率	85.10	9.29	84.00	100.00	71.00
平均值	84.04	8.54	83.80	99.40	70.20

可以看出，所有维度的平均分均超过 82 分（范围 82.47-85.10），表明模型在苏格拉底式对话任务中具备较高的综合能力。其中：成功拒绝率（85.10±9.29）和错误答案识别准确率（84.73±8.03）表现最优，反映模型对错误逻辑的检测和拒绝机制较为成熟；成功解释率（82.47±8.44）相对稍低，提示模型在复杂逻辑的清晰阐释能力上存在优化空间。

错误答案识别与成功拒绝率的正相关（二者均值 84.73 vs. 85.10，标准差相近），符合理论预期：准确识别错误是有效拒绝的前提，二者共同构成模型的"批判性能力"。总体质量与正确答案识别准确率的强关联（均值 83.40 vs. 84.50，中位数 84.50 vs. 84.00），说明用户对对话质量的评价更依赖模型对正确答案的稳定识别，而非单纯拒绝错误。

成功解释率的相对滞后（均值 82.47），可能与其依赖自然语言生成（NLG）的复杂性有关，需平衡逻辑严谨性与表达通俗性。

5.结论和讨论

为探讨如何通过多智能体大语言模型实现复杂推理数学题的苏格拉底式教学，提升复杂推理数学题的教学效果，本研究基于苏格拉底式教学理论与多智能体协作框架，构建了由教师 and 教学主管多智能体协同的对话生成系统。通过对 30 道高等数学题目进行多轮苏格拉底式对话模拟，并采用定性分析与多维定量评价相结合的方法，该方案能够通过引导式提问有效促进学生的自主思考，并帮助学生识别并纠正推理错误。多智能体结合苏格拉底教学法优点在于发挥苏格拉底教学法启发性、个性化教学的同时，极大程度地降低幻觉现象，如编造题目等。研究发现虽然模型在错误检测与逻辑拒绝方面表现优异（成功拒绝率 85.10±9.29，错误答案识别准确率 84.73±8.03），但在成功解释率相对滞后（82.47±8.44），且各维度标准差较高（7.66-9.29），反映模型对题目特征的敏感性及稳定性不足。

综上所述，本研究提出的多智能体苏格拉底式教学框架通过动态对话引导与错误根源挖掘，提升了学习者的主动思考与问题解决能力。该方案不仅为高等数学教育提供了可扩展的智能辅导工具，其模块化设计（如教学主管的监督机制、学生认知状态模拟）还可推广至物理、化学等需复杂推理的学科教学中。

本研究也存在一些局限性。首先，样本规模限制方面，30 道题目可能无法覆盖复杂推理数学问题的全领域特征，需扩展至更大规模、多类型题库。其次，评分主观性方面，人工模拟评分需通过多评委一致性检验（如 Cohen's Kappa）以提高效度。最后，因果关系待验证方面，需通过回归分析或结构方程模型（SEM）进一步量化各维度间的相互作用。

未来研究方向包括：1) 覆盖更多数学分支与题型，以验证该方法在各类数学问题中的普适性；2) 引入多模态交互机制，优化公式识别与可视化表达，降低因 Latex 转换误差导致的逻辑断裂风险；3) 量化教学策略的因果效应，通过结构方程模型（SEM）解析各评价维度间的相互作用路径；4) 增强解释可解释性，结合分步推理（Chain-of-Thought）与外部知识库验证，平衡逻辑严谨性与表达通俗性；5) 此外，需通过多评委一致性检验（如 Cohen's Kappa）降低评分主观性，并探索动态置信度阈值调整策略，以精细化模型的拒绝与解释机制。

参考文献

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Leoni Aleman, F., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). GPT-4 technical report.
- Adıgüzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*.
- Al-Hossami, E., Bunesco, R., Teehan, R., Powell, L., Mahajan, K., & Dorodchi, M. (2023). Socratic questioning of novice debuggers: A benchmark dataset and preliminary evaluations. In *Proceedings of BEA* (pp. 709–726).
- Alshaikh, Z., Tamang, L. J., & Rus, V. (2020). Experiments with a Socratic intelligent tutoring system for source code understanding. In *FLAIRS*.
- Briars, D. J., & Larkin, J. H. (1984). An integrated model of skill in solving elementary word problems. *Cognition and Instruction*, 1(3), 245–296.
- Chen, E., Huang, R., Chen, H.-S., Tseng, Y.-H., & Li, L.-Y. (2023). GPTutor: A ChatGPT-powered programming tool for code explanation. In *International Conference on Artificial Intelligence in Education* (pp. 321–327). Springer.
- Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 439–477.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems.
- Dan, Y., Lei, Z., Gu, Y., Li, Y., Yin, J., Lin, J., Ye, L., Tie, Z., Zhou, Y., Wang, Y., et al. (2023). EduChat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*.
- Ji, T., Lyu, C., Wang, L., Yu, D., Shi, S., Zhang, Z., & Tu, Z. (2023). Document-level machine translation with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Ju, L., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- Kushman, N., Artzi, Y., Zettlemoyer, L., & Barzilay, R. (2014). Learning to automatically solve algebra word problems. In *Proceedings of ACL* (pp. 271–281). <https://doi.org/10.3115/v1/P14-1026>
- Ma, H., Wang, C., Zhu, H., Yang, S., Zhang, X., & Zhang, X. (2024). Enhancing cognitive diagnosis using un-interacted exercises: A collaboration-aware mixed sampling approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38, 8877–8885.
- Paul, R., & Elder, L. (2007). Critical thinking: The art of Socratic questioning. *Journal of Developmental Education*, 31(1), 36.
- Qi, J., Xu, Z., Shen, Y., Liu, M., Jin, D., Wang, Q., & Huang, L. (2023). The art of Socratic questioning: Recursive thinking with large language models. In *Proceedings of EMNLP* (pp. 4177–4199).
- Shridhar, K., Macina, J., El-Assady, M., Sinha, T., Kapur, M., & Sachan, M. (2022). Automatic generation of Socratic subquestions for teaching math word problems.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- Wang, Y., Liu, X., & Shi, S. (2017). Deep neural solver for math word problems. In *Proceedings of EMNLP* (pp. 845–854).
- Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023). Document-level machine translation with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Yue, X., Qu, X., Zhang, G., Fu, Y., Huang, W., Sun, H., Su, Y., & Chen, W. (2023). Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

- Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., & Liu, W. (2023). MetaMath: Bootstrap your own mathematical questions for large language models. arXiv preprint arXiv:2309.12284.
- Zhao, W., Shang, M., Liu, Y., Wang, L., & Liu, J. (2020). Ape210k: A large-scale and template-rich dataset of math word problems. arXiv preprint arXiv:2009.11506.
- Zhou, L., Dai, S., & Chen, L. (2015). Learn to solve algebra word problems using quadratic programming. In Proceedings of EMNLP (pp. 817–822).