从答案到问题：EQGBench 用于评估大型语言模型的教育问题生成

Chengliang Zhou[1], Mei Wang[1], Ting Zhang[1], Qiannan Zhu[1], Jian Li[1], Hua Huang[1],
[1]School of Artificial Intelligence, Beijing Normal University

Abstract

大型语言模型（LLMs）在数学问题解决中展示了显著的能力。然而，从提供答案到生成高质量教育问题的转换呈现出重大挑战，这些挑战尚未深入研究。为推进教育问题生成（EQG）并帮助 LLMs 生成具有教育意义和教育效果的问题，我们引入了 EQGBench，这是一个专门为评估 LLMs 在中文 EQG 中的表现而设计的综合基准。EQGBench 建立了一个五维度的评估框架，支持数据集中包含 900 个评估样本，涵盖了三门基础中学学科：数学、物理和化学。该数据集结合了用户查询，具有不同的知识点、难度梯度和问题类型规格，以模拟真实的教育场景。通过对 46 个主流大型模型的系统评估，我们揭示了在生成反映教育价值并培养学生综合能力的问题上，仍有很大的发展空间。

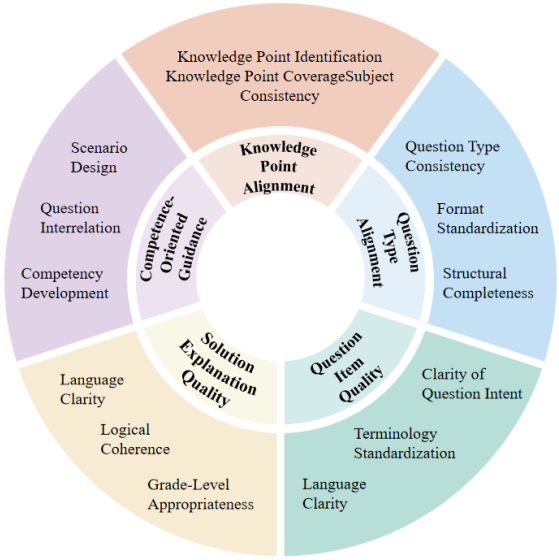

Figure 1: EGQBench 评估维度的设计及其相应的详细指标。

1 引言

从 GPT-3 的出现到 ChatGPT 和 GPT-4 的最新突破，大型语言模型（LLMs）在理解复杂查询和生成类似人类的响应方面展示了非凡的能力，特别是在数学问题解决的领域。然而，当我们从答案转向问题时，一个根本性的变化出现了：这些擅长提供解决方案的强大模型，能否掌握在教育背景下生成问题这一更具挑战性的任务？

在教育领域，生成高质量问题的能力是有效教学和学习的基石。虽然近年来出现了各种自动题目生成 (AQG) 方法 (Wang et al., 2020; Cho et al., 2019; Mulla and Gharpure, 2023)，但现有的方法主要关注从预定答案及其相应的上下文信息中导出问题，而非解决教育问题生成 (EQG) 所固有的独特需求。这一区别至关重要：EQG 强调基于特定的教学需求和学习目标生成问题。此外，EQG 不仅仅局限于表面的事实回忆，而是要培养更高层次的认知能力，包括概念理解、推理能力、解决问题的能力，从而对问题生成系统提出了更严格的要求。

为了促进 EQG 的快速发展，全面的评估基准至关重要。然而，许多现有的自动化问题生成（AQG）基准依赖于 BLEU 和 ROUGE 等基于 n 元组的指标。这样的评估模式与教育的目标基本上是不一致的。这些指标奖励的是表层的词汇相似性，但教育性问题的价值不在于其措辞，而在于其引发的认知过程。例如，这些指标无法区分简单的事实回忆问题和需要多步骤推理、概念应用或高阶思维的复杂问题。一个有效的教育问题会引导学生通过特定的问题解决路径，这是文本相似性算法完全看不见的维度。因此，依赖这些指标妨碍了生成在教学和学习上真正有效的问题的进展。

为了解决这一关键差距，我们引入了 EQG-Bench，一个专门设计的全面基准，用于评估模型生成高质量中文教育问题的能力。EQG-Bench 由一个精心策划的评估样本数据集支持，共有 900 个评估样本，涵盖数学、物理和化学三个基础中学学科，每个学科均匀分布 300 个样本。该数据集包含了不同知识点、难度梯度和问题类型规格的多样化用户查询，以真实模拟现实教育场景。此外，它提供了一个多维度评估框架，深度对齐教育目标，涵盖知

识点对齐、问题类型对齐、问题项质量、解决方案解释质量以及一个关键的以能力为导向的指导维度。通过将问题质量转化为一系列可量化和可解释的评估维度，我们对模型从“答案提供者”到“问题创造者”的转变过程进行了细粒度分析，探索它们的优势和局限性。

使用 EQGBench，我们对 46 个主流 LLM 进行了全面评估，包括来自 ChatGPT、DeepSeek 和 GLM 系列的模型。我们的实验结果表明，这些模型在基础理解任务上跨学科差异很小，而参数数量较大的模型在要求高阶逻辑推理的任务中具有明显优势。一个重要发现是，面向能力的指导在所有模型中都是一个显著的弱点。

本文的主要贡献如下：

1. 我们构建了一个高质量的 EQG 数据集，该数据集模拟了现实场景，涵盖了初中三门核心学科：数学、物理和化学。

2. 我们设计了一个五维的评估框架，以全面衡量所生成问题的内容质量和教育价值。

3. 我们对包括 ChatGPT、DeepSeek 和 GLM 在内的 46 种主流 LLMs 进行了系统评估。通过人工研究，我们验证了 EQGBench 在评估问题生成方面的科学严谨性和实际用途。

2 相关工作

2.1 问题生成

早期关于问题生成 (QG) 的研究主要依赖于基于模板的方法和神经序列到序列 (Seq2Seq) 模型。基于模板的方法使用知识点填充预定义的句子结构，但它们灵活性差且生成的内容单调 (Ali et al., 2010; Mitkov et al., 2003; Heilman and Smith, 2010; Mostow and Chen, 2009) 。虽然神经 Seq2Seq 模型能够从给定的上下文中生成相关问题，但它们在创造力和理解复杂指令方面的能力有限 (Zhou et al., 2018; Zhao et al., 2018; Dong et al., 2019; Cao et al., 2020) 。通过利用其强大的零样本和少样本能力，LLMs 现在可以通过精心设计的提示生成符合用户需求的问题，提供高灵活性和多功能性 (Maity et al., 2025; Maity and Deroy, 2024) 。

尽管这些技术上取得了显著进步，基于 LLM 的问题生成的评估仍然滞后。最近出现了一些中文 LLM 评估基准，例如高考基准 (Zhang et al., 2024) ，它使用全国高考题目来评估问题解决能力；C-EVAL (Huang et al., 2023) ，一个综合的中文语言评估套件；CMMLU (Li et al., 2024) ，一个多学科评估基准；以及 FinEval (Guo et al., 2024) ，一个金融知识的评估。然而，这些工作的共同点是它们专注于评估模型的知识储备和问题解决能力。它们未能评估模型在特定教学目标下生成符合课程标准的富有创意和见解的问题的能力。

2.2 问题评价

问题评价主要分为人工评价和自动化评价。人工评价需要专家，如具有专业知识的教师，提供跨越多个维度的综合分数。例如，一些研究使用众包工人在 1 到 5 的范围内为问题评分 (Du and Cardie, 2017, 2018) 。同样，MATHWELL (Christ et al., 2024) 是最近提出的框架，用以指导人工标注。然而，人工评价成本高、耗时长、难以扩展，其结果还可能受到评价者的主观性影响，使其不适用于快速评价大量模型。传统的自动化指标如 BLEU (Papineni et al., 2002) 、ROUGE (Lin, 2004) 和 METEOR (Banerjee and Lavie, 2005) 通过计算生成文本和参考文本之间的 n-gram 重叠来对问题进行评分。这些指标主要衡量表面级的文本相似性，无法有效评价问题的逻辑连贯性、可解性或教育价值。因此，对于需要高语义和逻辑准确度的问题生成任务来说，其效用极为有限。

使用大语言模型 (LLMs) 进行自动化评价的范式已成为新的研究热点。Zheng (Zheng et al., 2023) 和 Chiang (Chiang et al., 2023) 展示了使用 LLMs 作为评判者的可行性和可靠性。这种方法已被应用于作文评分 (Kim and Kim, 2024) 和数学答案评估 (Jiang et al., 2025; Urrutia and Araya, 2023) 。Wang (Wang et al., 2024) 提出了 PMAN 指标，该指标提示 LLM 回答其生成的问题以确定问题的有效性。在问题生成 (QG) 领域，一些研究也开始探索更深层次的评价维度。EduBench (Xu et al., 2025) 在更广泛的教育背景下评估模型，而 Dr.Academy (Chen et al., 2024) 则基于布鲁姆分类法评估问题生成能力。

尽管这些研究代表了积极的进展，但它们往往未能与实际的、真实世界的需求联系起来。相比之下，EQGBench 是一个全面的问题生成基准，直接与中学教育的核心教学原则和课程要求相联系。这种直接的对齐确保了我们的评估结果具有更大的实际意义和教学相关性。

3 数据集构建

在真实的教育环境中生成教育性问题是一个高度复杂且情境化的任务。它的要求远远超出了简单的知识点检索，对现有自动化系统的文本理解和生成能力提出了严峻挑战。目前，缺乏专门为这些复杂教育需求设计的系统化评价基准。

为了弥补这一差距，我们推出了 EQGBench，

一个全面的评估数据集，旨在系统地评估大型语言模型的教育问题生成能力。EQGBench 包含 900 个高质量的评估样本，这些样本均匀分布在数学、物理和化学领域。通过结构化的模板设计和动态信息填充，该数据集生成多样化的用户查询，这些查询在多个教育环境中（包括教师备课、学生个性化练习和家长指导）具有不同的知识点、难度梯度和问题类型规格。

3.1 模板构建

为了确保 EQGBench 的模板能够全面覆盖真实教学场景的需求，同时保证指令的质量和多样性，我们首先邀请了几位资深的中学教师为 EQG 设计了约 40 个初始指令，涵盖核心科目。基于这 40 个由人设计的指令，我们采用了参数化、重写和类比生成的三步过程，以创建一个更大且语言上更加多样化的提示集。

我们解构了初始说明，将学术阶段、科目、知识点、题目类型、难度等级和所需题目数量等核心要求抽象为参数化变量。这一过程形成了一组结构化的基本模板。

风格重写 我们利用了各种大语言模型，包括 Doubao、Qwen 和 DeepSeek，从多个角度（例如老师、学生和家长的角度）重写这些基础模板，以引入丰富的风格变化。

类比生成 我们使用重写的模板作为示例，以提示 LLMs 通过模仿生成更大规模的新提示模板语料库。

为了从结构化模板生成多样化的用户提示，我们采用了一种分层随机抽样策略。该方法根据预定义的分布动态填充模板中的多维参数，包括学术阶段、科目、问题数量、知识点、问题类型和难度。此过程生成了 900 条指令的最终评估数据集。生成的数据示例在表格 1 中显示。

数据补充的详细信息如下：

年级：统一设定为“初中”，以准确符合此义务教育阶段的教学要求。

主题：涵盖三个核心科学学科：“数学”、“物理”和“化学”，每个学科有 300 个样本。

问题数量：指令要求提出“单一”问题或“多个”问题（特别是 2 或 3 个）。对于每个主题，单一问题与多个问题的指令比例为 260:40。

知识：这些知识点来自每个学科的官方初中课程。这些知识库被组织成一个层次的、树状的结构，在其中概念通过逻辑层逐步详细化。具体而言，数学知识系统包括四个主要层级，最终达到终端知识点，而物理和化学系统则各有五个层级。

问题类型：指定了三种类型的问题：“单选题”、“填空题”和“问题”，分布比例为 4:3:3。

难度：设计了一个分层的难度分布。包括五种类型：“简单”，“中等”，“困难”，“由易到难”（渐进）和“由难到易”（倒退），按照 1:1:1:1:1 的比例均衡分配。

3.2 人工审核

为确保最终数据集的高质量，我们对所有成功生成的指令进行了彻底的人工审核。审核过程集中在三个关键方面：措辞的流畅性、用词的准确性和格式的一致性。此步骤旨在防止内容遗漏、生硬的措辞或格式错误，确保每个提示准确代表真实的用户查询情境。

4 评估指标设计

作为教育技术的一个关键应用，教育问题生成对大型语言模型（LLMs）构成了重大考验。这一能力的评估较为复杂，因为教学内容涉及多层次的知识系统，不同的教育背景对问题质量有不同的标准。因此，传统的评估方法难以全面衡量模型的问题生成能力。尽管最近的研究探索了使用 LLMs 进行评估的可行性 (Team et al., 2025; Yang et al., 2024; Ng and Fung, 2024)，一个专门用于问题生成能力的评估框架仍未充分开发。为了解决这一问题，我们构建了一个多维全面评估框架，以系统地测量 LLMs 在教育问题生成任务中的表现。该框架基于五个关键指标：知识点对齐、问题类型对齐、问题项质量、解决方案解释质量以及能力导向的指导。各维度的评分分为三个等级：优秀、良好和较差，分别对应分数 2、1 和 0。

这一维度评估生成的问题是否能准确识别和反映用户输入中指定的知识点，确保问题与指定主题一致。

问题类型对齐 (QT) 这个维度评估生成的问题类型（例如，选择题、填空题、问题）是否符合用户的选择，并遵循该类型的标准格式要求。例如，单项选择题应该包含四个选项；填空题应该提供下划线、括号或其他明确的答案指示标记；问题可以呈现为一个综合性问题，结合各种格式如选择或计算。

- 优秀：问题的类型与用户的规范一致，并且符合该类型的标准格式。
- 好：问题的类型大体上与用户的规格一致，但在细节或格式上存在小错误。
- 较差：问题类型与用户的说明完全不一致，或者格式过于混乱以至于无法识别。

题目项质量 (QQ) 这个维度评估生成的问题是否表达清晰，目标明确无歧义，使用标准术

	Template Sample	Specific Sample
Sample 1	I am self-studying 等级 { 主语 }, and I have currently reached the 知识 section. I would like a self-assessment exercise in the form of a { question_type }.	I am self-studying middle school mathematics, and I have currently reached the basics of rational numbers section. I would like a self-assessment exercise in the form of a solve-and-explain question.
Sample 2	I am a { 主题 } student teacher, and I need to design interactive board work for tomorrow's demo lesson. Please include a { question_type } in the 知识 section, with a difficulty level of { 困难 }, in accordance with the { 年级 } curriculum.	I am a mathematics student teacher, and I need to design interactive board work for tomorrow's demo lesson. Please include a single-choice in the maximum value problem of $y = ax^2 + bx + c$ section, with a difficulty level of easy, in accordance with the middle school curriculum.
Sample 3	My child is in { 等级 } this year, and he tells me that he can never understand the { 知识 } section in { 主题 }. Could you provide { num } 问题 _ 类型 questions for practice?	My child is in middle school this year, and he tells me that he can never understand the simplifying absolute values within a range section in mathematics. Could you provide 2 fill-in-the-blank questions for practice?
Sample 4	I need to consolidate the 知识 section in 年级 { 主体 }. Can you give me { 数 } questions with 难度 level?	I need to consolidate the real numbers section in middle school mathematics. Can you give me 3 questions with medium level?

Table 1: 来自 EQGbench 的示例数据。每种模板类型体现了用户场景及其具体需求，从教师、学生或家长的角度提出有针对性的请求，同时还有不具有特定角色的通用查询。

语，并且可以通过唯一或明确的答案来解决。这确保了学生能够准确理解问题的意图并完成任务。

- 优秀：问题清晰、简洁，易于学生理解。
- 好：问题的语言含糊不清，或者术语使用不当。
- 差：语言令人困惑或不清晰，存在显著问题，例如冗余、逻辑谬误或拼写错误。

此维度评估为生成的问题提供的解释的正确性、严谨性和完整性。它还要求解释中涉及的知识适合目标学术阶段的认知水平和课程要求，并且正确答案能够从解释中推导出。

能力导向指导（CG） 这个维度评估生成的问题是否整合或模拟了一个真实情境，包括但不限于文化背景、实际学科应用或现实生活情境。它衡量问题在指导学生应用知识和发展高阶能力方面的价值。

- 优秀：这个问题包含了一个丰富的背景情境，直接与解决问题相关。
- 较差：这个问题只是对知识点的纯粹抽象应用，缺乏任何背景设计。

对于开源模型的响应，我们在单个 NVIDIA A800 GPU 服务器上部署了参数规模较小的模型，每个服务器有 80GB 的内存。推理是在 Python 环境中使用 vLLM 框架进行的。对于闭源模型，我们使用了一台具有 16GB 内存的 NVIDIA 4060 GPU，并通过 OpenAI 库提供的 API 访问各种模型。

为了保持一致性，所有模型都设置了统一的模型参数。温度参数被设置为 0.6，以确保在生成问题时具有较高的随机性。最大输出长度被限制在 4096 个标记，以防止过度思考或输出过多。对于具有“思考”模式的模型，此功能被启用以促进更好的问题生成。

为了彻底检验不同架构和规模的 LLMs 在智能问题生成中的表现，我们选择了 46 个主流模型，包括 DeepSeek R1、ChatGPT、Qwen3 和 Gemini，参数规模从 70 亿到几百亿。这些模型包括专注于推理的专业模型和通用对话模型，使我们能够全面分析模型规模、架构和能力对问题生成质量的影响。用于响应生成的提示如图 **??** 所示。

反应评估 DeepSeek R1 被用作评估器模型。DeepSeek R1 具有深层语义理解、丰富的主题知识和敏锐的教育意图捕捉能力，非常适合用于评估生成问题的质量。我们将评估标准直接嵌入到评估提示中，以确保评估模型可以根据明确的指南对生成的问题进行评分并提供详细的评估报告。此外，为了提高评估结果的可靠性和稳定性，我们采用了多轮投票机制，以减少单次评估的随机误差。具体来说，每个样本经过三个独立的评分轮次，将众数选为最终分数。如果不存在众数，则通过算术平均作为最终分数。评估提示如图 2 所示。

4.1 实验结果

实验结果如表 2 所示。结果显示，在闭源通用模型中，Doubao-1.5-thinking-pro 在所有维度上表现最佳，在所有维度上得分均超过 1.9 分，除能力导向指导外，其得分相对较低。在所

Evaluation Prompt Design

System： You are an experienced middle school exam question designer with 20 years of expertise. Based on the following evaluation dimension, please strictly score the given question according to the scoring criteria, in combination with the user input and the generated question.

Query：
Dimension: Knowledge Point Alignment. This measures whether the generated question accurately matches and adequately covers the specified knowledge point.
Scoring Criteria:
0 points: The question item fails to correctly reflect the knowledge point. There is a significant mismatch between the knowledge point used in the question and the one specified by the user, or it is from a different subject.
1 point: The question item is generally relevant in topic but does not directly address or include the user-specified knowledge point.
2 points: The question item basically covers the user-specified knowledge point, with no significant omissions.
Example for 0 points:
Question Item: A city surveyed 1,000 residents to determine awareness of garbage sorting. Of those, 920 were aware. Based on this data, answer the following: 3. Estimate the margin of error for the city's garbage-sorting awareness rate at a 95% confidence level, rounded to two decimal places. Assume the population variance is unknown and estimate using sample variance.
Scoring Justification: “Confidence level” is not part of the middle school curriculum.
Example for 2 points:
User Request: Please create a multiple-choice question on addition and subtraction of polynomials at the middle school level.
Question Item:We define a linear equation in one variable $ax = b$ to be a “difference-solution equation” if its solution is $b - a$. For example, the solution of $2x = 4$ is 2, and $2 = 4 - 2$, so $2x = 4$ is a difference-solution equation.
Scoring Justification: Even though the concept is newly defined, the solution process involves addition and subtraction of polynomials, thus the specified knowledge point is covered.

User Input: {user_input}
Generated Question: {generate_question}

Note: As long as the question includes the specified knowledge point in any part, it is considered “basically covered” and earns 2 points. For example, if only one sub-question among several involves the knowledge point, or if only one option in a multiple-choice question does, it still counts as basic coverage.
Do not apply a lowest-score-first principle unless there are multiple sub-questions—in that case, the final score should be the lowest score among all the sub-questions.
Important: Only evaluate the content within the <question item> tags; ignore <solution explanation> and <answer>.
If the question is incomplete and cannot stand alone, the score is automatically 0 for this dimension.

Please output in the following format:
[Scoring Justification]: …<eoa>
[Score]: …<eoa>

Figure 2: 在 EGQBench 中使用的具有代表性的提示设计，用于评估模型生成的问题与知识点的对齐情况。

有三个科目中，该模型均排名前二。在开源通用模型中，DeepSeek-R1 在所有科目上表现出色。然而，Llama-3.1-8B-Instruct 在所有科目上表现最差，尤其在问题类型对齐、问题项质量和解决方案解释质量维度上尤为落后。在教育模型中，来自闭源设置的 Spark-X1 整体表现最佳，但仍落后于通用模型。开源的 educhat-base-002-13b 表现最差，所有维度得分都显著偏低。

4.1.1 量纲分析

模型在基础理解任务上的表现显示出整体稳定性，不同科目之间差异很小。在知识点对齐和题型对齐方面，闭源和开源模型总体得分都很高。例如，在表现优异的通用模型中，如 o4-mini、QwQ-32B 和 Doubao-1.5 系列，三个科目的得分都接近满分。在教育模型中，这两个维度的表现也普遍优于其他三个。这表明主流的大型语言模型具备很强的识别和映射问题基本结构及其核心评估点的能力。

在要求更高推理和逻辑能力的任务中，闭源模型和更大的参数开源模型表现出优势。在题目质量和解答解释质量维度上，表现出明显的分层，以 Doubao-1.5-thinking-pro 和 DeepSeek-R1 这样的模型获得了超过 1.9 的分数，而表现最差的模型分数则不超过 0.8。同时，在教育模型中，得分最高的模型相比于顶级通用模型仍显不足。一个显著的趋势是，由于数学领域有丰富的训练数据可用，模型在数学方面的表现更佳，往往在数学成绩上高于物理和化学。这表明了通用模型在推理方面的优势，而教育模型尽管更注重教育场景，在复杂认知任务上仍然存在困难。

在所有模型中，能力导向指导维度是最弱的。数学的得分尤其低，而物理和化学则表现相对较好。这表明模型在理解题目设计背后的教育意图方面仍然缺乏强大的能力，特别是在数学领域。几乎没有模型在数学中得分超过 0.3，而表现最好的物理和化学模型得分达到 0.9。大多数模型在这些维度上的得分在 0.7 左右，反映了模型在生成上下文相关问题的能力方面的当前差距，尤其是像数学这样更抽象的学科。

在整体表现方面排名前三的模型是 Doubao-1.5-thinking-pro、Gemini-2.5-flash-preview 和 o4-mini，总分超过 8。排名后三的模型是 educhat-base-002-13b、Llama-3.1-8B-Instruct 和 Spark-Lite，得分分别为 3.14、3.83 和 4.22。QwQ-32B 在知识点对齐方面获得满分，而 Doubao-1.5-pro-32k 和 Doubao-1.5-vision-pro 在问题类型对齐方面得分最高，DeepSeek-R1 在问题项质量方面获得了最高分 1.97。

排名前三的模型是 Doubao-1.5-thinking-

Model	Mathematics					Physics					Chemistry				
	KP	QT	QQ	SQ	CG	KP	QT	QQ	SQ	CG	KP	QT	QQ	SQ	CG
Qwen3-235B-A22B	1.98	1.68	1.72	1.75	0.23	1.99	1.75	1.57	1.64	0.86	2.00	1.72	1.64	1.69	0.71
Qwen3-8B	1.99	1.80	1.81	1.79	0.22	1.97	1.91	1.74	1.58	0.71	1.99	1.89	1.68	1.51	0.54
Qwen3-32B	1.98	1.95	1.87	1.84	0.23	2.00	1.96	1.73	1.69	0.97	1.99	1.97	1.87	1.74	0.69
QwQ-32B	2.00	1.90	1.89	1.87	0.21	2.00	1.94	1.80	1.80	0.92	1.98	1.95	1.85	1.81	0.65
Qwen2.5-72B-Instruct	1.95	1.80	1.84	1.74	0.18	1.97	1.87	1.72	1.57	0.65	1.96	1.92	1.79	1.63	0.60
DeepSeek-V3	1.97	1.89	1.79	1.76	0.20	1.98	1.92	1.69	1.67	0.93	1.98	1.98	1.87	1.79	0.64
DeepSeek-R1	1.98	1.95	1.95	1.96	0.17	1.99	1.99	1.95	1.94	0.77	1.99	1.93	1.91	1.94	0.73
GLM-Z1-32B	1.98	1.77	1.69	1.55	0.28	2.00	1.84	1.53	1.45	1.13	1.97	1.82	1.63	1.37	0.74
GLM-4-32B	1.96	1.66	1.67	1.64	0.19	1.97	1.74	1.59	1.62	0.78	1.99	1.80	1.67	1.59	0.47
GLM-Z1-9B	1.98	1.21	1.57	1.53	0.23	1.94	1.37	1.25	1.10	0.76	1.94	1.16	0.94	0.94	0.53
Yi-34B-Chat	1.79	1.08	1.18	0.60	0.15	1.86	1.49	1.33	0.95	0.58	1.95	1.54	1.28	0.85	0.23
internlm3-8b-instruct	1.72	0.75	1.24	1.13	0.24	1.79	1.09	1.20	1.11	0.73	1.84	1.31	1.15	1.11	0.52
Moonlight-16B-A3B-Instruct	1.61	1.29	1.42	1.11	0.25	1.65	1.43	1.19	0.83	0.63	1.71	1.65	1.38	0.90	0.56
Llama-4-Scout-17B-16E-Instruct	1.96	1.24	1.38	1.35	0.17	1.92	1.30	0.98	1.09	0.76	1.92	0.71	0.85	0.93	0.50
Llama-3.3-70B-Instruct	1.78	1.78	1.46	1.19	0.19	1.83	1.68	1.41	1.11	0.77	1.91	1.71	1.39	1.14	0.51
Llama-3.1-405B-Instruct	1.82	1.50	1.54	1.47	0.24	1.93	1.80	1.59	1.22	1.10	1.89	1.74	1.39	0.89	0.64
Llama-3.1-8B-Instruct	1.52	0.84	0.81	0.42	0.24	1.55	0.91	0.67	0.29	0.38	1.78	0.61	0.53	0.39	0.11
gemma-3-27b-it	1.89	1.86	1.64	1.59	0.17	1.92	1.88	1.42	1.14	0.90	1.94	1.83	1.50	1.09	0.71
gemma-3-12b-it	1.87	1.23	1.39	1.39	0.23	1.86	1.54	1.24	1.00	1.06	1.91	1.78	1.30	0.81	0.73
Mistral-Small-3.1-24B-Instruct	1.80	1.52	1.56	1.35	0.17	1.90	1.60	1.34	1.21	0.73	1.89	1.73	1.51	1.17	0.45
Mistral-Large-Instruct	1.86	1.63	1.68	1.57	0.17	1.90	1.81	1.69	1.42	0.57	1.97	1.89	1.75	1.49	0.32
Phi-4	1.85	1.86	1.61	1.49	0.14	1.91	1.88	1.58	1.26	0.63	1.91	1.88	1.60	1.22	0.37
doubao-1-5-thinking-pro	1.99	1.96	1.96	1.97	0.25	2.00	1.98	1.89	1.96	1.20	2.00	1.93	1.91	1.98	0.75
doubao-1.5-vision-pro	1.98	1.98	1.90	1.90	0.18	1.98	1.98	1.79	1.78	0.81	1.99	1.95	1.89	1.88	0.61
doubao-1-5-pro-32k	1.99	1.98	1.92	1.88	0.19	1.98	1.98	1.80	1.83	0.75	1.99	2.00	1.93	1.91	0.59
glm-4-Plus	1.89	1.69	1.65	1.56	0.19	1.83	1.77	1.56	1.31	1.18	1.91	1.93	1.57	1.17	0.77
glm-z1-air	1.96	1.80	1.73	1.60	0.30	1.96	1.89	1.53	1.29	1.31	1.95	1.85	1.58	1.08	0.74
qwen-plus-latest	1.99	1.87	1.90	1.87	0.21	1.98	1.96	1.82	1.83	0.87	1.99	1.97	1.87	1.86	0.70
qwen-max-latest	1.96	1.92	1.75	1.75	0.21	1.98	1.95	1.88	1.76	0.86	1.98	1.98	1.89	1.82	0.58
qwq-plus-latest	1.98	1.76	1.79	1.80	0.23	2.00	1.67	1.57	1.73	0.85	2.00	1.74	1.60	1.67	0.67
o4-mini	1.98	1.94	1.97	1.94	0.19	1.96	1.94	1.80	1.76	0.81	1.97	1.92	1.84	1.73	0.62
o3-mini	1.98	1.93	1.97	1.98	0.15	1.99	1.95	1.89	1.89	0.79	2.00	1.95	1.87	1.82	0.71
gpt-4.1	1.94	1.94	1.71	1.69	0.18	1.98	1.95	1.83	1.76	0.88	1.99	1.96	1.81	1.73	0.47
GPT-4o	1.96	1.82	1.76	1.72	0.21	1.99	1.90	1.80	1.75	0.85	1.99	1.95	1.80	1.69	0.49
Claude 3.7 Sonnet	1.87	1.73	1.65	1.52	0.19	1.93	1.88	1.46	1.29	1.16	1.97	1.96	1.52	1.07	0.83
Claude 3.5 Haiku	1.86	1.74	1.51	1.27	0.21	1.85	1.81	1.42	0.96	1.17	1.87	1.93	1.42	0.87	0.91
Gemini 2.5 Pro Preview	1.98	1.93	1.92	1.93	0.16	1.98	1.95	1.91	1.90	1.00	1.99	1.95	1.94	1.90	0.92
Gemini 2.5 Flash Preview	2.00	1.95	1.96	1.96	0.22	1.99	1.96	1.93	1.94	0.87	2.00	1.96	1.92	1.97	0.67
Moonshot-v1-32K	1.87	1.43	1.45	1.37	0.31	1.91	1.71	1.49	1.18	1.01	1.95	1.90	1.56	1.33	0.72
Hunyuan-Large	1.94	1.91	1.88	1.78	0.23	1.92	1.90	1.67	1.55	0.81	1.93	1.81	1.80	1.68	0.57
Yi-Lightning	1.91	1.89	1.75	1.57	0.14	1.98	1.95	1.71	1.67	0.95	1.97	1.96	1.80	1.65	0.70
educhat-base-002-13b	1.09	0.68	0.75	0.45	0.18	1.25	0.81	0.63	0.22	0.19	1.30	0.73	0.55	0.16	0.13
educhat-sft-002-13b	1.89	1.64	1.45	0.58	0.42	1.94	1.63	1.03	0.25	0.66	1.88	1.58	0.85	0.16	0.42
Confucius-o1	1.92	1.58	1.74	1.60	0.25	1.96	1.61	1.47	1.32	0.87	1.95	1.80	1.67	1.45	0.78
Spark-X1	1.99	1.63	1.70	1.76	0.22	2.00	1.71	1.50	1.56	1.13	1.98	1.67	1.57	1.64	0.74
Spark-lite	1.39	1.21	1.04	0.45	0.13	1.61	1.43	1.11	0.45	0.39	1.72	1.66	0.96	0.33	0.35

Table 2: 模型评分是使用 DeepSeek-R1 评估的。每个维度中的最高分以粗体显示，评估维度的全名在第 4.1 节中提供。

pro、Gemini-2.5-pro-preview 和 Gemini-2.5-flash-preview，得分分别为 9.02、8.74 和 8.69。Doubao-1.5-thinking-pro 在知识点对齐方面获得满分，而 DeepSeek-R1 在问题类型对齐和问题项质量方面排名第一，得分分别为 1.99 和 1.95。针对能力导向指导的最高得分为 1.31，由 glm-z1-air 获得。

前三个模型是 Gemini-2.5-pro-preview、Doubao-1.5-thinking-pro 和 Gemini-2.5-flash-preview，它们的得分分别为 8.71、8.58 和 8.51。Doubao-1.5-thinking-pro、o3-mini 和 Qwen3-235B-A22B 在知识点对齐方面获得了满分，而 Doubao-1.5-pro-32k 在问题类型对齐方面得分最高。Gemini-2.5-pro-preview 在问题项质量上得分最高，为 1.94。

虽然主要公司的顶尖模型在各学科中表现相似，但差异主要体现在能力导向的指导维度。这表明，尽管顶尖模型在中学各学科中的整体表现没有显著差异，但生成具有语境相关性和应用驱动的问题的能力仍然是一个关键的改进领域。

鉴于作为评估者的大型语言模型固有的不确定性和随机性，我们引入了专家人工评估来验证自动评估算法的有效性和可靠性。我们从数学评估集随机选择了 100 个测试样本，并获得 GLM-Z1-9B、Spark-X1 和 o4-mini 的结果。邀请了六位有经验的中学数学教师按照与自动评估相同的维度和标准进行独立评估。

为了评估人工评估者与自动评估者之间的一致性，我们采用了两种不同的评估方法。首先，我们通过直接计算人工评估者评分与自动评估者评分之间的数值差异，进行评分级别一致性评估（SC）。其次，我们通过分析人工评估者的排名与自动评估者对三种模型的排名之间的相关性，并使用斯皮尔曼等级相关系数，进行排名级别一致性评估（RC）。SC 和 RC 可以表述为：

$$SC_j = 1 - \frac{1}{K}\sum_{k=1}^{K} \left(\frac{1}{N}\sum_{i=1}^{N} \left| S^{model_k}(Q_i, D_j) - S^{human_k}(Q_i, D_j) \right| \right) \quad (1)$$

$$RC_j = 1 - \frac{6\sum_{K=1}^{K} d_K^2}{K\left(K^2-1\right)} \quad (2)$$

其中 $S^{human_k}(Q_i, D_j)$ 表示人工评估者针对 i -th 问题的 k -th 模型在 j -th 维度上的评分，$N = 100$ 是样本总数。$K = 3$ 是模型数量。d 表示在特定维度上给定模型的排名差异。

得分一致性结果显示，DeepSeek-R1 在所有维度上与人工评分达到超过 88 % 的一致性，其中在问题类型对齐方面的一致性最高，达到 97 %。$SRCC$ 结果表明大多数维度具有完美的相关性 1，其中解决方案解释质量维度的相关性为 0.5。这表明我们自动化评估框架与专家人工评估高度一致，确认所提出框架的有效性和可靠性。

	KP	QT	QQ	SQ	CG
SC	0.9650	0.9733	0.8983	0.8850	0.9197
RC	1.0000	1.0000	1.0000	0.5000	1.0000

Table 3: 通过计算 LLM 和人类评分获得不同维度的 SC 和 RC 分数。较高的 SC 表示模型评分与人类评分之间有更强的一致性，而较高的 RC 值则意味着两个排名系统之间的对齐更好。

5 结论

本文介绍了 EQGBench，这是一个用于评估大型语言模型（LLMs）教育问题生成能力的基准。EQGBench 包含一个高质量的数据集，包括数学、物理和化学领域的 900 个结构化指令，反映了多样的、真实世界的用户需求。其核心是一个多维框架，与教学目标对齐，对模型进行五个关键指标的评估。该框架的自动化流程显示出高度的可靠性和一致性，并已通过专家教师的评估验证。

我们对 46 个主流 LLM 的评估表明，虽然领先的模型具备强大的基础能力，但它们在生成具有深刻教育意图的问题时遇到困难。我们相信 EQGBench 是学术界的宝贵资源，将引导未来 LLM 在教育领域的优化。

References

Husam Ali, Yllias Chali, and Sadid A Hasan. 2010. Automatic question generation from sentences. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 213–218.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. 2020. Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1011–1020, New York, NY, USA. Association for Computing Machinery.

Yuyan Chen, Chenwei Wu, Songzhou Yan, Panjun Liu, Haoyu Zhou, and Yanghua Xiao. 2024. Dr.academy: A benchmark for evaluating questioning capability in education for large language models. *Preprint*, arXiv:2408.10947.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. *Preprint*, arXiv:1909.01953.

Bryan R Christ, Jonathan Kropko, and Thomas Hartvigsen. 2024. Mathwell: Generating educational math word problems using teacher annotations. *Preprint*, arXiv:2402.15861.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. *Unified language model pre-training for natural language understanding and generation*. Curran Associates Inc., Red Hook, NY, USA.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Copenhagen, Denmark. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. *Preprint*, arXiv:1805.05942.

Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. 2024. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *Preprint*, arXiv:2308.09975.

Michael Heilman and Noah A Smith. 2010. Good question! statistical ranking for question generation. In *Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pages 62991–63010. Curran Associates, Inc.

Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2025. Llms can find mathematical reasoning mistakes by pedagogical chain-of-thought. *Preprint*, arXiv:2405.06705.

Seungyoon Kim and Seungone Kim. 2024. Can language models evaluate human written text? case study on korean student writing for education. *Preprint*, arXiv:2407.17022.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Subhankar Maity and Aniket Deroy. 2024. The future of learning in the age of generative ai: Automated question generation and assessment with large language models. *Preprint*, arXiv:2410.09576.

Subhankar Maity, Aniket Deroy, and Sudeshna Sarkar. 2025. Can large language models meet the challenge of generating school-level questions? *Computers and Education: Artificial Intelligence*, 8:100370.

Ruslan Mitkov and 1 others. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing*, pages 17–22.

Jack Mostow and Wei Chen. 2009. Generating instruction automatically for the reading strategy of self-questioning. In *Artificial Intelligence in Education*, pages 465–472. IOS Press.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Prog. in Artif. Intell.*, 12(1):1–32.

Chee Ng and Yuen Fung. 2024. Educational personalized learning path planning with large language models. *Preprint*, arXiv:2407.11773.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, and 78 others. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Preprint*, arXiv:2502.14739.

Felipe Urrutia and Roberto Araya. 2023. Automatically detecting incoherent written math answers of fourth-graders. *Systems*, 11(7).

Zhen Wang, Siwei Rao, Jie Zhang, Zhen Qin, Guangjian Tian, and Jun Wang. 2020. Diversify question generation with continuous content selectors and question type modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2134–2143, Online. Association for Computational Linguistics.

Zifan Wang, Kotaro Funakoshi, and Manabu Okumura. 2024. Automatic answerability evaluation for question generation. *Preprint*, arXiv:2309.12546.

Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Yang Gao, and Heyan Huang. 2025. Edubench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios. *Preprint*, arXiv:2505.16160.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S. Bernstein, and John Mitchell. 2024. Social skill training with large language models. *Preprint*, arXiv:2404.04204.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2024. Evaluating the performance of large language models on gaokao benchmark. *Preprint*, arXiv:2305.12474.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3901–3910.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing*, pages 662–671, Cham. Springer International Publishing.