大语言模型在国际中文阅读自动出题中的效能评估
(Evaluating the Effectiveness of Large Language Models for Automatic Question Generation in International Chinese Reading)

景宏伟
(Jing, Hongwei)
北京语言大学
(Beijing Language and Culture University)
jhw080266@163.com

徐娟
(Xu, Juan)
北京语言大学
(Beijing Language and Culture University)
xujuan@blcu.edu.cn

摘要：随着国际中文教育数字化转型的持续推进，传统人工命题模式在效率、成本与规模化应用方面的瓶颈日益凸显。在此背景下，以大模型为代表的人工智能技术，为自动命题提供了新的技术路径。本研究以 HSK6 级阅读理解题为研究对象，系统评估大模型在自动命题任务中的实际效能。研究选取四种大模型，结合提示工程开展实验，涵盖指令大模型、推理大模型以及经 LoRA 微调后的垂直大模型。并从语言流畅度、内容准确性、题目复杂度、选项干扰性、答案唯一性、题型多样性六个维度，辅以 BLEU、ROUGE、Distinct 等机器指标，对生成题目进行综合评估。研究表明：大模型生成的题目与人工命题具有较高的相似性，但在难度控制、可回答性等方面尚存在不稳定性，需经人工审核修订后方可用于教学；在模型对比中，推理大模型整体表现更优。基于此，本研究进一步提出相应的使用建议，以优化题目生成过程，推动人机协同命题模式的发展。

Abstract: With the continued advancement of the digital transformation of international Chinese language education, the traditional manual approach to test item development has increasingly encountered bottlenecks in terms of efficiency, cost, and scalability. Against this backdrop, artificial intelligence technologies, particularly large language models (LLMs), have opened up new possibilities for automated test item generation. This study focuses on HSK Level 6 reading comprehension items and systematically evaluates the practical effectiveness of LLMs in automatic item generation. Four LLMs were selected for experimentation using prompt engineering, including instruction-tuned models, reasoning-oriented models, and a domain-specific model fine-tuned with Low-Rank Adaptation (LoRA). The generated items were comprehensively evaluated across six dimensions: linguistic fluency, content accuracy, item complexity, distractor quality, answer uniqueness, and item-type diversity. In addition, machine-based evaluation metrics, including BLEU, ROUGE, and Distinct, were employed to provide complementary assessments. The results indicate that the items generated by LLMs exhibit a high degree of similarity to those developed

by human experts. However, the models still demonstrate instability in controlling item difficulty and ensuring answerability, suggesting that human review and revision remain necessary before the generated items can be used in instructional settings. Among the models evaluated, reasoning-oriented LLMs achieved the best overall performance. Based on these findings, this study further proposes practical recommendations for optimizing the item generation process and advancing a human–AI collaborative approach to test development.

关键词: 国际中文教育，生成式人工智能，大语言模型，题目自动生成，HSK

Keywords: International Chinese Language Education, Generative Artificial Intelligence, Large Language Model, Automatic question generation, HSK

1. 引言

题目编制是语言测试与教学评估体系中的核心环节，其质量直接影响测评结果的有效性与教学反馈的准确性。长期以来，国际中文教育领域主要依赖人工方式进行题目设计与开发。然而，随着全球中文学习需求的持续增长，截至 2025 年 HSK 全球累计考生规模已超过 850 万人（郁云峰等，2025），传统人工命题在效率、成本控制及大规模题库建设等方面的局限性日益凸显，已难以满足快速增长的测评需求。

此外，HSK 3.0 考试体系的推出（曹贤文等，2025），对题目数量、质量及更新速度提出了更高要求，尤其是在标准化与大规模应用场景下，传统人工命题模式在短时间内难以高效产出高质量试题，进而可能对 HSK 3.0 的推广与实施效果产生一定制约。近年来，以大模型为代表的人工智能技术快速发展，在文本生成、逻辑推理等方面展现出显著优势，为自动化题目生成提供了新的技术路径与实现可能。

本研究依托 HSK3.0 考试体系，以 HSK6 级阅读理解题为研究切入点，向大模型输入阅读材料以生成相应题目。同时，从多维度对生成题目进行人工评估，并结合机器评价指标，以提升评估结果的客观性与全面性。本研究旨在探究大模型在国际中文教育自动命题任务中的能力边界，为国际中文教师基于大模型开展自动命题提供实践参考。

2. 研究综述

2.1 大模型赋能国际中文教育资源研发

近年来，以大模型为代表的生成式人工智能技术正持续推动国际中文教育领域的数字化转型，为教学资源的智能化开发与个性化建设注入了新动能。目前，相关应用主要集中于智能化内容生成方向，具体体现为：在文本资源方面，支持分级阅读材料的自动生成（韩欣欣等，2025）、个性化阅读材料的生成（侯泽煜、徐娟，2025）；在写作教学方面，能够辅助开发智慧化写作资源如智慧教材、范文语料库等（马瑞祯、徐娟，2024）；在课程资源建设方面，可助力高效生成结构化的中文微课内容（李嘉仪、徐娟，2025）。这些实践不仅提升了资源开发的效率与多样性，也为实现精准化、个性化的国际中文教学提供了有力支持。而在国际中文教育题目资源建设中，刘玉屏等（2025）探索了生成式人工智能赋能 HSK 模拟试题的编写。

2.2 题目自动生成

题目自动生成（Automatic Question Generation，AQG)技术属于文本生成的一项子任务，伴随着自然语言处理技术的发展而演变。早期 AQG 主要采用基于规则的方法，例如 Mitkov 等人（2003）提出的基于句法模式匹配的框架，通过语法分析和模板填充生成问题。这类方法能确保生成问题的规范性，但受限于模板库，往往存在多样性不足的问题。进入统计机器学习阶段，通过引入概率模型提升问题生成的灵活性，如 Heilman 等（2010）提出了一种基于规则和统计排序的 AQG 方法，其核心思想是“过度生成再排序”，通过规则生成大量候选问题，再利用统计模型对这些问题进行排序以筛选出最优结果。但规则构建依然依赖人工，成本较高。

随着神经网络的发展，深度学习方法显著推动了 AQG 向语义理解和自然表达的方向发展，显著提升了题目生成的语义连贯性与多样性。Jiang 和 Lee（2017）将词嵌入模型应用于汉语名词多项选择题干扰项自动生成中，通过分布式表示计算词汇语义相似度，优化了干扰项设计。Du 等（2017）提出将基于注意力机制的序列到序列（Sequence-to-Sequence，Seq2Seq）模型应用于 AQG 任务，实现从文本中自动生成阅读理解题目。徐坚（2023）进一步提出融合门控循环单元与图注意力网络的增强型 Seq2Seq 模型，通过答案引导的图注意力机制捕捉文章内部依赖关系，并结合注意力机制与指针网络，提升了所生成题目的语义关联与答案确定性。然而，该阶段仍存在逻辑一致性不足、认知层次较浅等问题。

Transformer 架构（Vaswani et al，2017）凭借并行计算和自注意力机制有效解决了长距离依赖问题，为 AQG 提供了更好的语义理解和表达能力。基于 Transformer 架构的预训练大语言模型，如 BERT（Devlin et al，2019）、T5（Raffel et al，2020）和 GPT 系列（Brown et al，2020）等，通过在海量文本数据上预训练，学习了丰富的语言表示和知识，获得了强大的语言理解和生成能力，为 AQG 提供了新的方向。陈欣等（2024）结合提示工程构建了一种基于大模型的试题自动生成路径。来雨轩等（2024）为激发大模型在 AQG 任务上的潜力，提出了将大模型和检索增强技术

相结合的生成方法。聚焦在国际中文教育领域，有学者尝试利用大模型进行阅读测试题的研发（王鸿滨、吕海辉，2025；王亚敏等，2025）。目前，大模型在国际中文教育 AQG 任务上的能力边界尚有待进一步验证。

总体而言，大模型在自动出题任务上展现出来较大的潜力，但尚存在不足，本研究尝试探索以下工作：

1）通过高效参数微调构造垂直领域大模型，验证该方法在国际中文教育自动出题任务上的表现；

2）从主客观相结合的多维角度，系统对比国内外不同类别大模型包括指令型、推理型以及垂直领域大模型在出题任务上的表现；

3）提出大模型在国际中文教育自动出题任务上的人机协同命题模式，为一线教师和命题专家提供参考。

3. 研究设计

为全面评测大模型在国际中文教育自动命题任务上的表现，本研究采用实验比较法，通过系统化的研究设计对大模型自动出题效能进行评测。其设计核心环节包括：首先基于模型代表性和中文处理能力的综合考量，选取典型大模型作为评测对象；其次以 HSK6 级真题为主要来源构建数据集；最后通过标准化的实验操作流程和科学的评价指标，确保实验数据的可靠性和可比性。以下将分别从评测模型选择、数据集构建、实验设计以及评估策略四个维度详细阐述研究设计。

3.1 模型选择

在模型的选择上，本研究参考了中文语言理解测评基准 SuperCLUE 榜单[1]，该榜单聚焦于通用大模型的综合性测评。选取了四款在 SuperCLUE 通用榜中排名靠前的大模型，具体模型信息如表 1 所示。

表 1　大模型具体信息

模型名称	发布机构	是否推理	属地	发布时间
Gemini-3-Pro-Preview	Google	是	海外	2025.11
DeepSeek-V3.2-Thinking	深度求索	是	国内	2025.12
Qwen-3-Max	阿里巴巴	否	国内	2025.11
Llama-4-Maverick-17B-128E-Instruct	Meta	否	海外	2025.11

选取的四个模型中包含了来自海内外的推理型、指令型大模型。其中 Gemini-3-Pro-Preview 是谷歌推出的多模态大模型，在数学推理、代码生成及跨模态理解方面突出，尤其擅长科学问答与复杂逻辑推理，属于推理大模型；DeepSeek-V3.2-Thinking

[1] 网址参见：https://superclueai.com/generalpage。

是深度求索推出的推理专项模型，通过思维链增强与自我反思机制，显著提升复杂推理、长文本分析与分步求解能力，属于推理大模型；Qwen-3-Max 是阿里通义千问高性能版本，在中文理解、长上下文处理与多语言任务中表现优异，适合长文档分析与生成，属于指令大模型；Llama-4-Maverick-17B-128E-Instruct 是基于 Llama 4 的高效指令微调模型，擅长指令跟随与多轮对话，属于指令大模型。此外，还有一类垂直大模型，即针对特定任务进行专门训练与优化的大模型。由于目前没有公开的可用于自动命题的国际中文教育领域垂直大模型，本研究采用微调技术构建用于出题的垂直大模型。考虑到算力资源、硬件设施、数据集规模等因素，本研究选取了参数量较小的大模型 Qwen2.5-7B 作为微调的基座模型。

3.2 数据集构建

在微调模型的训练集构建中，数据主要来源于 2016 年至今的 HSK6 级阅读真题以及模拟题，覆盖多种体裁、多种题型，以确保数据的权威性、规范性与教学适配性，同时可以将模型生成的题目与真题进行对比，以探究大模型在 AQG 任务上的能力边界。数据集具体构建流程为：首先，进行数据的采集和筛选，初步搜集到了 680 余篇真题材料，均为 PDF 格式；其次，进行文本提取，经 OCR 技术将其转换为可编辑文本、并剔除冗余信息，通过关键词定位阅读理解题模块获取阅读材料及其对应题目；然后，进行语料清洗与去重，修正 OCR 识别残留的错别字与语句不通问题并确保与原始阅读材料段落划分一致，同时采用余弦相似度算法计算文本间相似度，进行语料去重，确保无重复阅读材料，最终确认 640 篇阅读材料；最后，调整数据格式，将数据整理为“阅读材料—题目”的结构化数据格式，以适配模型的训练。此外，考虑到现有数据集规模过小，故另外选取了 600 篇 HSK6 级模拟题，最终训练集包含 1240 条数据。需要说明的是，模拟题在题型结构与知识点覆盖上参考 HSK6 级考试要求设计，但其来源于非官方命题体系，与真题在命题规范性方面仍存在一定差异，本研究在模型训练过程中将其作为与真题同分布的近似数据进行使用，以增强模型对 HSK6 级阅读理解题型结构的学习能力。此外，在测试集的构建中，本研究另外选取了 30 篇 HSK6 级阅读真题，利用测试集中的阅读材料，对不同大模型进行自动出题能力的评测。需要说明的是，测试集与训练集中的阅读材料完全独立，没有重复。

3.3 实验设计

3.3.1 提示词设计

在利用通用大模型进行问题生成时，本研究通过编写提示词的方法实现。提示词设计是大模型应用中的关键环节，对于充分发挥大模型能力至关重要，它通过规范化、结构化的指令引导模型理解任务意图、约束输出范围，从而显著提升生成内容的准确性、相关性和可控性。提示工程（Prompt Engineering）是一种通过设计、实验和优化输入来引导模型生成高质量、准确和有针对性的输出的技术（Dong et al，2024），其中输入的格式一般称作提示模板，组织各种提示信息的方式称为提示策略或提示方法，其中常用的提示策略如表 2 所示。

表 2 常用的提示策略

提示策略	描述
零样本提示（Zero-Shot）	大模型在没有任何任务示例的情况下，仅依据自然语言指令执行任务
少样本提示（Few-Shot）	通过提供少量示例引导模型执行任务
思维链提示（Chain-of-Thought，CoT）	在解决复杂推理问题时，要求模型将中间推理步骤显式地输出，鼓励模型展示推理步骤以提升复杂问题解答的准确性
角色扮演提示（Role-Playing）	通过指令为模型赋予特定角色或身份以控制输出风格与内容

目前，在提示词的设计上已有很多通用法则和实践经验。为提升大模型的输出质量，学界探索了多种结构化提示框架以优化提示工程效果，例如具有代表性的 ICIO 框架、CLEAR 框架等[2]。此外，针对不同应用情境，也形成了相应的提示框架。在教育领域，CRISPE 框架（王华树、谢斐，2024）在实践中得到广泛应用，并被证实能够有效提升模型回答的质量。基于上述考虑，本研究选用 CRISPE 框架作为提示词设计的模板，具体提示词设计如表 3 所示。

表 3 题目生成提示词框架

组成部分	示例内容
角色（Capacity and Role）	你是汉语水平考试（HSK）的命题专家
背景（Insight）	中文学习者在应对 HSK6 级阅读时，常对长难句理解、隐含意图推断、文化负载词把握及篇章逻辑衔接等方面存在困难。
任务（Statement）	请基于提供的阅读材料，设计四道高质量的选择题。要求每道题包含四个选项，只有一个答案正确，所有题目必须基于材料内容，难度符合 HSK6 级水平。
格式（Personality）	题目格式示例： 1.老总当初为什么要留这个年轻人？ A 公司急需人员 B 客户欣赏年轻人 C 相信自己没有看错人 D 年轻人有丰富的工作经验 答案：C

为验证该提示词的有效性，本研究首先进行了小规模的预实验，发现尽管初始提示词设定了基本框架，但模型生成的题目在考查点分布、难度控制及选项设计方面存在改进空间。具体而言，生成的部分题目未能精准对应 HSK6 级的核心能力要求，部分干扰项的干扰性不足或偏离原文逻辑，题目难度呈现不稳定现象。为了系

[2] 网址参见：https://developer.aliyun.com/article/1490356。

统提升生成题目的质量与规范性，对初始提示词进行了结构化迭代与优化，优化后的提示词框架如表 4 所示。

表 4 优化后的题目生成提示词框架

组成部分	示例内容
角色（Capacity and Role）	你是精通 HSK6 级的资深命题专家，深谙考试大纲，可以精准把握命题要求。
背景（Insight）	在 HSK6 级阅读命题中，应重点考查学习者对长难句的理解、对隐含观点或态度的推断、对特定语境下词语的理解，以及对篇章整体逻辑与主旨的把握等。
任务（Statement）	请严格依据提供的阅读材料，设计四道单项选择题。要求每道题包含四个选项，只有一个答案正确，要求题型具有多样性，包括细节题、推理判断题、主旨大意题、词义题等。题目应体现 HSK6 级应有的认知复杂度，避免仅进行原文词句的简单匹配。
格式（Personality）	请严格按照以下题目格式示例输出： 1.老总当初为什么要留这个年轻人？ A 公司急需人员 B 客户欣赏年轻人 C 相信自己没有看错人 D 年轻人有丰富的工作经验 答案：C

3.3.2 模型参数设置

此外，为批量生成题目，本研究通过调用大模型 API 的方式实现自动出题，在不同模型的参数设置上参考了官方的默认参数，主要包括采样多样性参数（Tempreture）和采样范围调节参数（top_p）。Tempreture 是一个用于调节模型输出概率分布“平滑度”的超参数，它通过在对数概率（logits）被转换为概率（softmax）之前，对其进行缩放，从而控制生成过程的随机性。在计算下一个词的概率时，模型原始的 logits 向量会除以 Temperature 的值 T，见公式 1 所示，其中，w_i 为词表中的第 i 个词元（token），z_i 是模型为词表中第 i 个词元输出的原始 logit 值，V 是词表大小。Top-p 是一种通过设定概率阈值，在文本生成的每一步动态筛选出最小的高概率候选词集合进行采样，以在保证连贯性的前提下控制输出多样性的自适应方法。

$$P(w_i) = \frac{\exp(z_i/T)}{\sum_{j=1}^{V} \exp(z_j/T)} \quad \text{公式（1）}$$

通过合理调整 Tempreture 和 Top-p，可以引导大模型在生成文本的“创造性”和“可控性”之间找到最佳平衡。在本研究中，各个模型的参数设置参考了官方的默认参数设置以及在实际应用中的效果，具体参数信息如表 5 所示。

表 5 模型参数设置

模型	Tempreture	Top-p
Gemini-3-Pro-Preview	1.0	0.95
DeepSeek-V3.2-Thinking	1.0	0.95
Qwen-3-Max	0.8	0.9
Llama-4-Maverick-17B-128E-Instruct	0.8	0.9

3.3.3 垂直领域大模型构建

在构造自动出题的垂直大模型时，本研究采用有监督微调（Supervised Fine-Tuning，SFT），在模型训练方式的选择上，采取了 LoRA 微调（Low-Rank Adaptation）方法。LoRA 是一种针对大型预训练语言模型的高效微调技术，它旨在解决全参数微调所带来的计算和存储成本问题，其核心思想是冻结预训练模型的原始参数，并通过引入少量可训练的低秩矩阵来模拟参数更新。这样在微调过程中，只需要优化这些低秩矩阵的参数，而不需要修改原始模型的参数，从而大大减少了需要训练的参数量（Manakul et al，2023）。本研究选取了 LLaMA-Factory 作为微调工具，这是一个专为大模型微调而设计的低代码训练框架，它提供了一套完整的工具和接口，以简化和加速大模型的训练、微调和部署过程。微调时的超参数配置如表 6 所示。

表 6 微调超参数设置

参数	取值
训练轮数	3
学习率	3e-4
批处理大小	16
权重衰减系数	0.01
学习率调整策略	Linear
LoRa 秩值	8

3.4 评估策略

在评估大模型生成题目的质量时，为确保评估的全面性和有效性，本研究采取客观指标与主观评价相结合的评估方法。

3.4.1 客观指标

题目自动生成作为文本生成的一项子任务，因此本研究参考了文本生成任务中常用的一些机器指标来评估题目质量。选取了 BLEU（Bilingual Evaluation Understudy）（Papineni et al，2002）、ROUGE（Recall-Oriented Understudy for Gisting Evaluation）（Lin，2004）以及 Distinct（Li，2016）三个指标，取值范围均为 0 至 1，通常用百分比表示，其中利用 BLEU、ROUGE 指标来判断大模型生成题目与参考题目（真题）的表面相似性，利用 Distinct 指标来判断题目的多样性。BLEU 主要用来评估自动出题和参考题目之间的 n-gram 的重叠程度（即相似度），本研究分别将 n 设置为

1、2、3、4，然后求四个结果的均值，计算公式见式（2），其中 BP 为惩罚因子，避免因为题目过短而给出过高分数，计算公式见式（3），lr 表示最短的参考题目长度，lc 为模型生成的题目长度；ROUGE 运用 n-gram 上的召回率（Recall）来衡量自动生成的题目与参考题目之间的相似度，计算公式见式（4），n 通常取值为 1、2 和 L，本研究使用 ROUGE-L 值（最长公共子序列的匹配度）来进行评估；Distinct 是评估文本生成多样性的指标，通过统计生成文本中不重复的 n-gram 的比例来衡量词汇多样性，包括 macro-distinct 和 micro-distinct，其中 macro-distinct 关注单个文本，而 micro-distinct 关注生成的全部文本，本研究使用 micro-distinct 来评估生成题目的多样性，其中 n 设置为 2，计算公式见式（5）。

$$\mathrm{BlEU} = \mathrm{BP} \times exp\left(\sum_{\mathrm{n}=1}^{N} \mathrm{W_n} \times \log \mathrm{P_n}\right) \quad \text{公式（2）}$$

$$\mathrm{BP} = \begin{cases} 1，\ \mathrm{lr} < \mathrm{lc} \\ exp（1 - \mathrm{lr}/\mathrm{lc}），\ \mathrm{lr} \geq \mathrm{lc} \end{cases} \quad \text{公式（3）}$$

$$\mathrm{ROUGE-N} = \frac{\sum_{\mathrm{S}} \sum_{gramN} count_{\mathrm{match}}（gram_N）}{\sum_{\mathrm{S}} \sum_{gramN} count（gram_N）} \times 100\% \quad \text{公式（4）}$$

$$\mathrm{Distinct-N} = \frac{count(uningram)}{count(ngram)} \quad \text{公式（5）}$$

3.4.2 主观评估

在进行人工评价时，参考韩雨婷等（2025）梳理的基于大模型的题目自动生成系统专家审核维度体系，并结合对 HSK 试题的特点分析，从“语言流畅度”“内容准确性”“题目复杂度”“选项干扰性”“答案唯一性”“题型多样性”六个维度出发对生成的题目进行系统评估。其中，“题型多样性”指标是针对同一篇阅读材料所生成的四道题目，在考查形式（如细节题、主旨题、推断题、词义理解题等）上的分布，其余五个维度均针对单道题目进行评价。

在评估人员选择方面，为提升评分效度，所有参与的评分人员均需具备 HSK 命题经验、国际中文教育相关背景或丰富的一线教学经历。本研究共邀请 4 位有经验的命题员参与，包括 1 位北京语言大学国际中文学院讲师（教龄 3 年，多次参与 HSK 命题），以及 3 位北京语言大学评价院语言测试方向博士（均具备 HSK6 级教学经验并多次参与汉语考试命题）。

在具体评估流程上，首先针对测试集中的 30 篇阅读材料，使用不同大模型逐篇生成题目，并与真题进行混编。混编后形成的评估题库共包含 180 篇阅读材料，每篇材料附有四道题目，题目来源涵盖真题与不同模型生成的题目。该题库以盲评形式分发给评估人员，即人工在评分时不知题目具体来源，仅依据上述六个维度进行逐题评估。

在评分方式上，“语言流畅度”“内容准确性”“题目复杂度”“选项干扰性”“题型多样性”五个指标均采用李克特五级量表进行评分，分值越高代表评估人员对该项指标越认同，其中 1 表示很不满意、2 表示不满意、3 表示一般满意、4 表示比较满意、5 表示很满意。其中，“答案唯一性”为定类变量，分为“唯一”与“不唯一”两类，数值 1 代表答案唯一，数值 0 代表答案不唯一包括不存在正确答案和存在多个正确答案两种情况。

4. 实验结果与分析

4.1 生成题目示例

以测试集中的一篇阅读材料为例，不同模型生成的题目对比如表 7 所示，该材料来源于 H61330[3]真题中的第一篇阅读理解题。

表 7 不同模型生成题目对比示例

题目来源	题干	选项
Gemini-3-Pro-Preview	根据文中提到的历史演变，唐朝时面条被称作：	A 煮饼 B 汤饼 C 水引 D 冷淘
	关于山西面食，下列哪项正确？	A 只有蒸和煮两种主要制法 B 有据可查的面食种类超过 280 种 C 所有的面食都统称为“记心火烧” D 家庭主妇通常只会做三四种面食
	孩子入学第一天吃“记心火烧”，寄托了父母怎样的愿望？	A 岁月延绵，年年有余 B 健康长寿，一生平安 C 能够增长学问和心眼 D 能够学会制作各种面食
	最适合做上文标题的是：	A 面食的制作工艺 B 山西面食文化 C 丰富多彩的面食名称 D 中外面食的差异
DeepSeek-V3.2-Thinking	根据材料，面条在唐朝被称作什么？	A 煮饼 B 汤饼 C 水引 D 冷淘
	材料中提到，山西有据可查的面食有多少种？	A 数十种 B 约 200 种 C280 种 D 超过 300 种
	关于山西面食所承载的文化寓意，下列哪一项说法正确？	A 过年吃“接年面”寓意长寿 B 孩子上学吃“记心火烧”是希望孩

[3] H61330 是 HSK6 级考试试卷的编号。

		子专心学习 C 过生日吃拉面是希望年年有余 D“接年面”是在生日时食用的
	外国友人评价“世界面食在中国，中国面食在山西”，旨在说明什么？	A 山西面食的历史最为悠久 B 山西面食的种类最为丰富 C 山西是面食文化的杰出代表 D 山西面食的烹饪方法最复杂
Qwen-3-Max	根据文章，唐朝时期面条的称呼是什么？	A 煮饼 B 汤饼 C 水引 D 冷淘
	文中提到“记心火烧”是在什么场合食用的？	A 过生日时 B 过年时 C 孩子第一天上学时 D 家人远行时
	从文中可以推断出，山西面食之所以种类繁多，主要是因为：	A 山西地理环境适合种植多种粮食作物 B 当地人对面食有深厚的情感和文化认同 C 外国游客的大量需求推动了面食创新 D 政府大力推广面食制作技艺
	文中“娇儿宠称多”这句话的意思最接近于：	A 孩子被宠爱时会得到很多昵称 B 被珍视的事物往往拥有多种称呼 C 面食名称复杂，令人难以记住 D 山西人喜欢给食物起新奇的名字
Llama-4-Maverick-17B-128E-Instruct	第 1 段中的画线词语“推陈出新”是什么意思？	A 翻新 B 更新 C 增加 D 更换
	下列哪一项是山西面食被誉为中国五大面食之一的代表？	A 拉面 B 刀削面 C 记心火烧 D 接年面
	孩子第一天去上学时，山西人会让孩子吃“记心火烧”，其寓意是？	A 希望孩子长寿安康 B 希望孩子多学知识 C 希望岁月延绵不绝 D 希望家中年年有余
	下列哪一项最适合作为本文的标题？	A 山西面食的历史渊源 B 名扬海外的刀削面 C 面食之乡——山西 D 山西面食的制作方法
微调后的模型	山西的刀削面被誉为中国五大面食之一，据材料可知山西有据可查的面食有多少种？	A 数十种 B 两百种 C280 种 D 五百种
	关于山西面食，下列哪项正	A 制法单一

	确？	B 花样繁多 C 以牛肉为主 D 多吃有利于健康
	根据上文，山西人为什么喜欢把面食叫做“精神食粮”？	A 面食营养丰富 B 面食品种繁多 C 面食能治病 D 面食寄托了人们的感情
	下列对文章内容的理解，不正确的一项是？	A 山西刀削面是中国五大面食之一，享誉海外 B 山西面食的称谓变化与时间、地域因素有关 C 山西面食仅能作为充饥的食物，没有其他价值 D 外国友人对山西面食的地位给予了高度认可

4.2 题目质量评估

本研究从客观指标和人工评估两个方面对大模型生成的题目进行全面评价，以确保评估的全面性和准确性。

4.2.1 客观指标

BLEU、ROUGE-L 以及 Distinct 指标的计算结果如表 8 所示。从整体上看，不同模型生成题目的 BLEU 值和 ROUGE-L 值均不高，这主要是由于生成的题目均为选择题，且题干和选项内容均较短，经分词处理后在机器指标上的表现并不好。

表 8 客观机器指标

模型	BLEU	ROUGE-L	Distinct
Gemini-3-Pro-Preview	19.76%	27.42%	66.67%
DeepSeek-V3.2-Thinking	20.64%	27.26%	64.18%
Qwen-3-Max	15.48%	24.64%	65.45%
Llama-4-Maverick-17B-128E-Instruct	16.26%	22.78%	52.08%
微调后的模型	13.24%	20.36%	53.99%

从内容质量指标 BLEU、ROUGE-L 上来看，通用大模型整体优于特定的微调大模型。其中，Gemini-3-Pro-Preview 与 DeepSeek-V3.2-Thinking 表现最为突出，BLEU 分数分别达到 19.76%与 20.64%，ROUGE-L 分数均超过 27%，说明二者在词汇匹配与语义覆盖方面与参考题目具有相对较高的相似性，生成的题目在内容上更贴近人工命题风格；在多样性指标 Distinct 上，Gemini-3-Pro-Preview 与 Qwen-3-Max 分别取得 66.67%与 65.45%的最高值，表明其生成题目的用词变化丰富，避免了重复与模板化表达。而微调后的模型虽然在 Distinct 上略高于 Llama-4-Maverick，但仍显著低于其它模型，反映出其生成文本的多样性相对有限。总的来说，通用大模型特别

是 Gemini-3-Pro-Preview、DeepSeek-V3.2-Thinking 两个推理大模型在题目生成的内容相关性与语言多样性方面均表现更佳；而专门微调的模型并未显示出预期优势。

4.2.2 人工评估

人工评分结果如表 9 所示，以真题得分为基准，重点考察各模型与真题的接近程度，差值越小代表模型表现越佳，越接近人工命题水平。其中括号中的数据表示模型在不同维度上得分与真题的差距，高于真题为“+”，低于为“-”。

表 9 评分结果

题目来源	语言流畅度	内容准确性	题目复杂度	选项干扰性	题型多样性
真题	4.58	4.65	3.96	3.92	3.98
Gemini-3-Pro-Preview	**4.56（-0.02）**	4.55（-0.10）	3.26（-0.70）	**3.55（-0.37）**	**4.02（+0.04）**
DeepSeek-V3.2-Thinking	4.61（+0.03）	**4.62（-0.03）**	**3.34（-0.62）**	3.46（-0.46）	4.10（（+0.12）
Qwen-3-Max	4.50（-0.08）	4.38（-0.27）	3.08（-0.88）	3.28（-0.64）	3.92（-0.06）
Llama-4-Maverick-17B-128E-Instruct	4.55（-0.03）	4.24（-0.41）	3.21（-0.75）	3.15（-0.77）	3.74（-0.24）
微调后的模型	4.48（-0.10）	4.08（-0.57）	3.28（-0.68）	2.90（-1.02）	3.25（-0.73）

从整体上看，所有模型在语言流畅度与内容准确性两个基础维度上表现最佳，与真题得分差距极小，表明大模型在语言规范性与内容忠实性上较为接近人工命题水平。然而，在体现命题专业重要能力的题目复杂度与选项干扰性维度上，所有模型均与真题存在显著差距，显示出大模型在高阶认知考查与精细选项设计方面存在明显短板。具体而言：

在语言流畅度上，大模型生成题目的流畅度与人类较为接近，在大部分题目中评估人员没有感受到人与机器的明显差异，显示了大模型强大的自然语言生成能力，生成的文本具有较好的拟人度。

在内容准确性上，DeepSeek-V3.2-Thinking 和 Gemini-3-Pro-Preview 的表现最好，与真题的差距更小，展现出其强大的上下文理解能力，生成题目与原文具有较高的一致性。

在题目复杂度上，大模型生成题目与真题存在明显差距，不能很好地控制题目难度，其中 DeepSeek-V3.2-Thinking 的表现与真题差距最小，经评估人员反馈，大模型生成的部分题目存在难度过大的情况，选项中出现难度过大的成语（如“病入膏肓”）以及难以分辨的近义词（如“翻新”和“更新”）。

在选项干扰性上，Gemini-3-Pro-Preview 相对表现最好，部分题目错误选项的干扰性较小，不能有效设计出基于典型错误、具有合理迷惑性的选项，如词语辨析、固定词语搭配等。

在题型多样性上，除微调后的模型表现较差外，其它模型表现与真题较为接近，覆盖了细节题、推理题、词义题、主旨大意题等多种题型。

不同大模型生成题目的答案唯一性占比如表 10 所示。易知，通用大模型生成题目具备较好的可回答性，答案唯一性占比均超过 90%；而经微调后的模型在该指标上表现显著落后，所生成的题目存在“无正确选项”和“多个正确选项”的情况，不符合单项选择题的基本设计要求。这一结果说明，答案设计的逻辑自洽性是影响大模型命题质量的关键因素，后续可通过加强答案定位验证机制、严格检验选项唯一性等方法，进一步提升生成题目的可靠性与可用性。

表 10 答案唯一性指标

题目来源	答案唯一	答案不唯一
Gemini-3-Pro-Preview	92.26%	7.74%
DeepSeek-V3.2-Thinking	92.15%	7.85%
Qwen-3-Max	93.64%	6.36%
Llama-4-Maverick-17B-128E-Instruct	91.35%	8.65%
微调后的模型	82.47%	17.53%

此外，从不同类别大模型的表现差异来看：首先，以 DeepSeek-V3.2-Thinking、Gemini-3-Pro-Preview 为代表的推理型大模型综合表现最优。这很可能得益于其内部的思维链推理机制，使其在题目生成过程中能够更好地模拟人工命题的认知流程，逐步理解文本、定位考查点、构思干扰项并确保答案唯一性，从而在内容准确性、选项干扰性等重要维度上更接近真题；其次，以 Qwen-3-Max 为代表的指令型大模型在语言流畅度与题型多样性等维度上表现良好，说明其能够较好遵循生成指令与格式要求。然而，其在题目复杂度与选项干扰性等需深层文本理解与逻辑设计的指标上略逊于推理型模型；最后，本研究所采用的微调垂直大模型整体表现不佳，可能在当前训练数据规模与微调策略下，模型未能充分融合领域知识并保持原有生成能力，可见，在优质训练数据受限的情况下，微调并非是有效提升大模型执行具体教学任务能力的手段。

5. 讨论

总的来说，大模型在题目生成任务上具有显著潜力，但仍在一定程度上存在题目难度控制、题目可回答性等问题，未来应建立人机交互的题目自动生成机制，重点发挥大模型在题目生成中的效率优势，并通过人工审核来提高生成题目的可靠性。为帮助教师减负增效，本研究提出以下两条使用建议。

5.1 优化提示设计与模型遴选，实现精准生成

教师应依据具体的教学目标与考查重点，采用结构化、精细化的提示设计并选择适配的大模型，以提升生成题目的质量与适用性。在提示设计中，应明确题目考查的认知层次、题型及难度要求，并可提供少量示例作为格式与风格的参考，从而有效引导模型输出。此外，可通过迭代优化提示内容，逐步形成稳定、高效的结构化提示模板。在模型选择上，可依据任务特性进行遴选：若侧重语言规范与格式准确性，可优先选用指令遵循能力强的模型如 Qwen-3-Max；若需加强题目的逻辑深度与高阶思维考查，则建议选用具有显式推理能力的模型如DeepSeek-V3.2-Thinking。通过结构化的提示设计与模型匹配，能够显著提升生成题目与教学情境的契合度与可用性。

5.2 建立“生成—审核—修订”的人机协同流程，实现闭环优化

教师可将大模型作为题目的初步批量生成工具，随后基于专业判断对生成结果进行审核。审核应重点关注题目与教学目标的契合度、难度是否合理、题目是否可回答。对于未达标的题目，教师可进行针对性修订或提供明确修改指令，重新投入生成环节。通过多次“生成—审核—修订”的迭代，形成持续优化的闭环流程。此机制既能发挥大模型在快速大量生成上的效率优势，又能确保最终题目经过严格的专业把关，从而在提升命题效率的同时保障题目的科学性与适用性。

6. 结语

本研究以 HSK6 级阅读理解题为例，探索了国内外不同类别大模型自动出题的效果，为大模型赋能教育领域自动出题的智能化转型提供实践参考。研究表明，大模型在自动出题任务上具有显著潜力，不同类别模型表现存在差异，其中推理大模型的整体表现更优，但在选项设计、难度控制、可回答性等方面与人工命题尚存在明显差距，无法直接用于教学实践，因此尚不能替代专业命题教师，而更适合作为辅助生成工具，在人工审核与修订的基础上投入教学使用。未来，可以进一步探索多模态大模型在图文题、听力题等多种题型上的应用效果，推动国际中文教育在资源建设上的数智化转型。

致谢：本文受教育部中外语言合作交流中心国际中文教育研究课题重大项目“面向国际中文教育的生成式人工智能（AIGC）应用研究”（24YH03A）；北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）“融合LLM和GraphRAG的自动命题方法研究”（26YCX040）的资助。徐娟为本文通讯作者。

参考文献

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A.,

Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

Cao, X. W., Feng, L. P., Wu, Z. W., Lu, G., Wang, J. M., Zhang, Y. L., Zhu, Y., Liang, X., & Zhang, X. S. (2025). Discussion on the new HSK 3.0 syllabus and exam promotion. *Journal of Yunnan Normal University (Teaching and Research of Chinese as a Foreign Language), 23*, 1-8.[曹贤文, 冯丽萍, 吴中伟, 路广, 王佶旻, 张艳莉, 朱勇, 梁霞, & 张新生. (2025). “HSK3.0 新考纲及考试推广”大家谈. *云南师范大学学报(对外汉语教学与研究版), 23*, 1-8.]

Chen, X., Li, M. R., Zhou, Y. Q., Zhou, T., & Zhang, F.(2024). Research on the automatic generation path of test questions based on large language models. *China Examinations, 2024*, 39-48. [陈欣, 李蜜如, 周悦琦, 周同, & 张峰. (2024). 基于大语言模型的试题自动生成路径研究. *中国考试, 2024*, 39-48.]

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics.

Dong, Q., Li, L., Dai, D., Xu, C., Zhu, Y., Sun, G., Sun, C., Jiang, S., Jia, Y., Sui, Z., & Chang, B. (2023). *A survey on in-context learning.* arXiv. https://doi.org/10.48550/arXiv.2301.00234

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1342-1352). Association for Computational Linguistics.

Han, X. X., Ma, R. L., & Xu, J. (2025). Exploring the technical path of DeepSeek empowering international Chinese teaching resource construction: A case study of graded reading text generation. *International Chinese Language Teaching Research, 2025*, 30-40. [韩欣欣, 马瑞祾, & 徐娟. (2025). DeepSeek 赋能国际中文教学资源建设的技术路径探索——以分级阅读文本生成为例. *国际汉语教学研究, 2025*, 30-40.]

Han, Y. T., Wang, W. X., Liu, H. Y., & You, X. F. (2025). Technological innovations and practical challenges in automatic question generation. *Advances in Psychological Science, 33*, 1766-1782. [韩雨婷, 王文轩, 刘红云, & 游晓锋. (2025). 题目自动生成的技术革新与现实挑战. *心理科学进展, 33*, 1766-1782.]

Heilman, M., & Smith, N. A. (2010). Good question! Statistical ranking for question generation. In *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 609-617). Association for Computational Linguistics.

Hou, Z. Y., & Xu, J. (2025). Research on personalized generation of international Chinese reading materials based on large language models. *International Chinese Language Education (Chinese and English), 10*, 32-44. [侯泽煜, & 徐娟. (2025).

基于大语言模型的国际中文阅读材料个性化生成研究. *国际中文教育(中英文), 10,* 32-44.]

Jiang, S., & Lee, J. S. Y. (2017). Distractor generation for Chinese fill-in-the-blank items. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 143-148). Association for Computational Linguistics.

Lai, Y. X., Wang, Y. D., & Wang, L. (2024). Subject test question generation method based on large language model and retrieval enhancement. *Journal of Chinese Information Processing, 38*, 148-158. [来雨轩, 王艺丹, & 王立. (2024). 基于大语言模型与检索增强的学科试题生成方法. *中文信息学报, 38*,148-158.]

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 110-119). Association for Computational Linguistics.

Li, J. Y., & Xu, J. (2025). Development and evaluation of international Chinese micro-courses based on generative artificial intelligence: A case study of elementary level sentence patterns. *Journal of Gannan Normal University, 46*, 53-61. [李嘉仪, & 徐娟. (2025). 基于生成式人工智能的国际中文微课开发与评价——以初等水平句式为例. *赣南师范大学学报, 46,* 53-61.]

Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop* (pp. 74-81). Association for Computational Linguistics.

Liu, Y. P., Ou, Z. G., & Wu, X. Q. (2025). Effectiveness evaluation of generative artificial intelligence empowering international Chinese teaching: A case study of instructional design, HSK mock test question writing, and essay scoring. *Journal of Ethnic Education Research, 36* 156-166. [刘玉屏, 欧志刚, & 武晓琴. (2025). 生成式人工智能赋能国际中文教学的效果测评——以教学设计、HSK 模拟试题编写及作文评分为例. *民族教育研究, 36,* 156-166.]

Ma, R. L., & Xu, J. (2024). Innovative research and development of intelligent teaching resources for international Chinese writing in the digital intelligence era. *International Chinese Language Teaching Research, 2024*, 13-23. [马瑞祾, & 徐娟. (2024). 数智时代国际中文写作智慧教学资源创新研发. *国际汉语教学研究, 2024,* 13-23.]

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). *SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models.* arXiv. https://doi.org/10.48550/arXiv.2303.08896

Mitkov, R., & Ha, L. A. (2003). Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing* (pp. 17-22). Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1-67.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30,* 5998-6008.

Wang, H. B., & Lyu, H. H. (2025). Research on automatic generation of Chinese reading test questions based on large language models. *International Chinese Language Teaching Research, 2025*, 41-54. [王鸿滨, & 吕海辉. (2025). 基于大语言模型的中文阅读测试题自动生成研究. *国际汉语教学研究, 2025,* 41-54.]

Wang, H. S., & Xie, F. (2024). Research on innovation of translation education practice model driven by large language model technology. *Chinese Translators Journal, 45*, 70-78. [王华树, & 谢斐. (2024). 大语言模型技术驱动下翻译教育实践模式创新研究. *中国翻译, 45,* 70-78.]

Wang, Y. M., Bin, S., & Zhao, Y. (2025). Research on the development and application of large language models empowering Chinese as a second language reading tests. *Language Teaching and Linguistic Studies, 2025*, 1-12. [王亚敏, 宾帅, & 赵杨. (2025). 大语言模型赋能中文二语阅读测试的研发与应用研究. *语言教学与研究, 2025,* 1-12.]

Xu, J. (2023). Automatic generation of reading comprehension questions supported by semantic graphs. *Journal of Intelligent Systems, 19*, 420-428. [徐坚. (2023). 语义图支持的阅读理解型问题的自动生成. *智能系统学报, 19,* 420-428.]

Yu, Y. F., Qian, X. H., Zhong, Y. H., Li, Y. M., Song, M. S., Xie, X. Q., Han, B. C., Ma, X. N., Gu, C. Y., Bai, L. M., Bai, J. H., Hao, Q. X.,Li, D. G., & Ru, S. (2025). HSK3.0 and the new ecology of international Chinese education: Standard innovation, technological empowerment and global development—A multi-perspective discussion. *International Chinese Language Education, 10*, 1-17. [郁云峰, 钱旭红, 钟英华, 李宇明, 宋明顺, 谢小庆, 韩宝成, 马西尼, 古川裕, 白罗米, 白建华, 郝清新, 李登贵, & 茹丝. (2025). “HSK3.0 与国际中文教育的新生态：标准革新、技术赋能与全球发展”多人谈. *国际中文教育(中英文), 10,* 1-17.]

