

生成式人工智能生成初中英语听力文本和试题能力探析

——以ChatGPT-4o和文心大模型-3.5为例

叶 丛

浙江师范大学外国语学院, 浙江 金华

收稿日期: 2025年1月16日; 录用日期: 2025年2月18日; 发布日期: 2025年2月25日

摘 要

本文对比了ChatGPT-4o和文心大模型-3.5生成初中英语听力文本和试题的能力, 以2023年杭州市中考英语听力真题为样本, 使用两种人工智能分别生成文本和试题。对文本和试题的分析结果表明, 在生成听力文本上, ChatGPT和文心大模型差距较小, 二者都能生成话题、词数、难度与样本文本基本一致的文本。在生成听力试题上, ChatGPT的表现好于文心大模型, 相比前者, 后者生成的试题出现了更多问题。笔者认为, 鉴于寻找或编写听力文本的困难, 英语教师在为学生编写低利害的练习题时, 使用生成式人工智能来生成听力文本是一个不错的选择, 但要生成高质量的试题, 则需反复修改, 以及需要更多人类智慧的参与。

关键词

人工智能, 听力理解, 试题开发

Investigating the Capability of Generative Artificial Intelligence for Generating English Listening Comprehension Texts and Items for Junior High School Students

—A Case Study of ChatGPT-4o and Large Model ERNIE-3.5

Cong Ye

College of Foreign Languages, Zhejiang Normal University, Jinhua Zhejiang

Received: Jan. 16th, 2025; accepted: Feb. 18th, 2025; published: Feb. 25th, 2025

Abstract

This paper compared the capability of ChatGPT-4o and Large Model ERNIE-3.5 to generate English listening comprehension texts and items for junior high school students, using the 2023 Hangzhou city Zhongkao listening texts and items as samples. Both artificial intelligence models were used to generate texts and items. The analysis results showed that in terms of generating listening texts, ChatGPT and Large Model ERNIE exhibited minimal differences, as both can produce texts with topics, word count, and difficulty similar to the sample texts. In generating listening items, ChatGPT performed better than Large Model ERNIE, with the latter producing more problematic items. The author argues that, given the difficulty of finding or creating listening texts, using generative AI to produce listening texts is a good option for English teachers when preparing low-stakes practice exercises for students. However, to generate high-quality test items, repeated revisions and more human engagement are necessary.

Keywords

Artificial Intelligence, Listening Comprehension, Test Development

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,我国高度重视教育信息化和学校教育数字化转型。党的二十大报告更明确强调“推进教育数字化”。教育数字化转型是指将传统的课堂教学模式向基于包括人工智能在内的信息技术数字化教学方式转变,以实现高效、便捷、可重复性高的教育服务[1],而推进人工智能在教育领域中的应用是教育数字化转型的重要路径[2]。近年来,基于大语言模型的生成式人工智能得到了迅猛发展,其中具有代表性的 ChatGPT (Chat Generative Pre-trained Transformer),一经问世就在多个领域产生巨大影响。ChatGPT 是由 OpenAI 开发的一种基于 GPT 模型的人工智能聊天系统,它通过使用大量的自然语言数据进行训练,能够理解和生成自然语言文本,模拟人类的对话。在教育领域,学者们探讨了 ChatGPT 展现出的巨大潜力。在教学方面,它可以生成教学素材、设计教学方案等;在学习方面,它可以提供知识问答,提供学习支架等;在评价方面,它可以生成测试题目,提供反馈建议等[3]。虽然 ChatGPT 可能带来教育诚信危机和伦理安全问题,但 ChatGPT 等新一代智能技术的深度应用将有望助力解决教育现代化进程中面临的重大问题,教师应积极探索与应用智能技术,从而优化教学[4]。与此同时,国内也开发了多种人工智能,如百度的文心一言,阿里云的通义千问,科大讯飞的星火认知大模型。这些国产人工智能在教育领域的应用潜力如何,能否平替 ChatGPT,目前还鲜有研究。

2. 文献回顾

自动化项目生成(Automatic Item Generation, AIG)的定义是使用计算机技术来大量生成试题,它最早可追溯到 20 世纪 70 年代,有着几十年的研究历史。与人工命题相比, AIG 的优点是省时省力。AIG 方法大致可分为基于模板的 AIG 和基于人工智能的 AIG [5]。在过去的 AIG 方法中,基于模板的 AIG 仍依赖专家编写试题模板,基于人工智能的 AIG 方法除了试题质量不如模板法生成的试题以外,还需要编程

知识,有一定门槛。作为最新的 AIG 方法,使用 ChatGPT 命题的优点在于其人机互动简单,只需输入提示词即可生成试题,因此适用于一线英语教师[6]。由于 ChatGPT 发布时间较近(2022 年 11 月),关于其应用于生成试题的研究数量仍有限,但这一领域正在获得越来越多的关注。以下是一些重要研究。

Aryadoust 等探索了 ChatGPT-4 生成不同难度听力测试的能力。通过提示词编写和微调,他们生成了学术、初级、中级、高级四种难度的听力文本和试题。研究发现虽然 ChatGPT-4 能够生成不同难度的听力文本,但试题的选项往往较长且语义相似。该研究说明了 ChatGPT-4 有降低测试开发成本的潜力,但也强调了需要专家审核以改进生成内容[7]。

Lin 和 Chen 探索了 ChatGPT 生成阅读理解多项选择题的能力,并通过心理测量分析和专家评审对其进行了评估。研究发现,ChatGPT 生成的试题与人工编写的试题在心理测量特性上相似,但有时选项过长且存在重叠。研究结果表明,尽管 ChatGPT 可以降低开发成本,但仍需仔细审核和优化,以确保试题质量[8]。

Shin 和 Lee 研究了 ChatGPT 生成二语测试材料的能力,并将其与韩国大学入学考试的测试材料进行了比较。对 50 名教师的调查表明,尽管 ChatGPT 生成的阅读材料有着相似的自然流畅的特点,但大学入学考试的试题质量更高、选项更具吸引力。该研究建议,ChatGPT 可以帮助英语教师,但需要教师进行人工改进[9]。

O 对比评估了 ChatGPT 生成的平行试题与人工编写的试题。通过专家评估和在大学生中进行试测,结果表明这两种形式的测试具有可比性,这说明 ChatGPT 可以帮助教师减少工作量和节省时间。然而,进一步的研究仍需进行,以优化 AI 生成的测试内容[6]。

在中国,关于 ChatGPT 生成试题的实证研究极为有限。据笔者所知,只有两项相关研究。

杨志明等基于中国英语能力等级量表、核心素养和高考评估体系的要求,使用 ChatGPT 生成了英语客观题和主观题。研究结果表明,ChatGPT 在命题方面具有很大的潜力,然而,生成的试题可能存在科学性错误,以及难易程度不太好把握。因此,ChatGPT 生成的试题应由专家团队把关,并且最好在使用前进行试测,以确保质量[10]。

梅凌宇等基于高中数学课程、高考评估体系、数学教科书、高考试卷和模拟试题,使用 ChatGPT 生成了数学试题。研究结果表明,ChatGPT 在命题方面具备一定的潜力,但其当前能力还不足以能够胜任高中数学试题的命制工作[11]。

综上所述,ChatGPT 在试题开发方面具有一定潜力。通过仔细编写提示词和反复修改生成内容,ChatGPT 可以生成与人工编写的试题质量相当的试题。然而,生成试题的质量仍需进行人工审核。尽管命题过程仍然需要人工参与,但这仍大大节省了命题的时间和精力。

关于国内生成式人工智能命题的研究极少,国内生成式人工智能是否能在命题方面平替 ChatGPT 仍有待研究。

3. 研究方法

本文使用 ChatGPT-4o 和文心大模型-3.5,使用同样的提示词使其生成初中英语听力文本和试题,然后进行对比分析。在生成文本时,为了使其生成与真题文本更相似的文本,笔者使用 2023 年杭州市中考听力真题文本作为样本,但使用样本和不使用样本生成的文本孰优孰劣目前尚无结论。本文使用两种人工智能分别生成 3 段小对话及其对应的 3 道试题,1 段长对话及其对应的 3 道试题,1 段独白及其对应的 4 道试题。在听力微技能上,高考考纲规定了英语听力应考查理解细节信息、理解主旨大意、作出推理判断、理解意图态度四种能力。在真题中,对于理解细节信息以外的能力考查较少,比例失衡,因此本文在参照真题考查的听力微技能的基础上,使用两种人工智能额外生成 1 道推理判断题、主旨大意题、意

图态度题。生成文本的提示词如下：

[样本文本]基于这段文本，生成一段初中英语听力短对话(或长对话、独白)，话题，词数，词汇语法难度与该文本一致。

提示词中包括样本文本，学段(初中)、文本类型、话题、词数、词汇语法难度等元素，由于使用了样本，因此通过“与该文本一致”这段提示词来规定生成文本的这几种特点。

生成试题的提示词如下：

[生成文本]基于生成的文本，生成一(多)道三选一选择题，考查理解细节信息(或其它听力微技能)的能力。题干和选项要简短，并对原文进行同义替换，选项长度、类型一致，干扰项要合理且具有迷惑性，且尽量使用文中已有信息命制。

提示词中包括题型，考查的听力微技能，以及英语听力单选题命题时应遵循的一些原则[12]。生成试题考查的听力微技能规定与原试题一致。

在生成完成后，对真题和人工智能生成的文本及试题进行对比分析。本文分析两种人工智能第一次生成的文本及试题，仅在生成文本长度不符时重新生成，不作其它修改。文本和试题的分析框架如下(表 1)：

Table 1. Analysis framework for texts and test items
表 1. 文本和试题分析框架

文本	话题
	长度
	难度
试题	题干和选项是否简短
	题干和选项是否对原文进行同义替换
	选项长度、类型是否一致
	干扰项是否使用文中信息命制
	考查的听力微技能

4. 研究结果与讨论

4.1. 文本

4.1.1. 话题

如表 2 所示，从话题上看，生成文本的话题与样本文本基本一致，符合要求。样本文本的短对话 1 的话题是购物时寻求帮助，ChatGPT 生成的短对话 1 去掉了购物的情境，仅为寻求帮助，因此归为日常活动的话题，文心大模型生成的短对话 1 则保留了样本文本的大部分内容，仅替换了部分词汇。样本文本的短对话 3 的话题是询问对方之前缺席的原因，原因是拜访祖父母，因此归为日常活动的话题，ChatGPT 生成的短对话 3 中说话人缺席的原因是生病在家休息，因此归为个人情况的话题，文心大模型生成的短对话 3 仅对样本文本进行了扩写，相似度很高。样本文本的独白话题是说话人介绍自己的写作经历，ChatGPT 生成的独白话题是说话人介绍自己练习钢琴和创作音乐的经历，文心大模型生成的独白话题与样本完全一致，并保留了样本文本的大部分信息，仅替换了表达方式，可看作是对样本文本的同义改写。总而言之，文心大模型生成的文本对样本文本改动相对较小，而 ChatGPT 改动相对较大。

Table 2. Analysis of text topics
表 2. 文本话题分析

文本话题	样本文本	ChatGPT	文心大模型
短对话 1	购物	日常活动	购物
短对话 2	学校生活	学校生活	学校生活
短对话 3	日常活动	个人情况	日常活动
长对话	计划与愿望	计划与愿望	计划与愿望
独白	个人情况	个人情况	个人情况

4.1.2. 长度

如表 3 所示，从长度上看，ChatGPT 和文心大模型生成的文本长度和样本文本基本一致，仅长对话长度超过样本文本，但仍处于真题词数范围之内，符合要求。其中，文心大模型倾向于生成词数远超过样本文本的文本，需要令其重新生成，ChatGPT 生成的短对话长度低于文心大模型，但二者生成的长对话和独白词数接近。

Table 3. Analysis of test length
表 3. 文本长度分析

文本话题	样本文本	ChatGPT	文心大模型
短对话 1	25 词	23 词	35 词
短对话 2	40 词	30 词	32 词
短对话 3	24 词	29 词	35 词
长对话	96 词	125 词	126 词
独白	192 词	194 词	180 词

4.1.3. 难度

本文根据弗莱士易读度(Flesch Reading Ease)衡量样本和生成文本的难度。弗莱士易读度是根据句子的字数和句子中含的音节数等计算的，数值越大，文章越容易读。由于对短对话计算弗莱士易读度会产生误差，因此将所有文本合在一起计算。从表 4 可以看出，三者文本易读度非常接近，说明三者文本难度基本没有差别，符合要求。

Table 4. Analysis of test difficulty
表 4. 文本难度分析

难度	样本文本	ChatGPT	文心大模型
总易读度	87.80	86.49	89.19

4.2. 试题

4.2.1. 命题原则

如表 5 所示，从题干长度看，真题题干长度一般为 7 词左右，最多不超过 11 词，ChatGPT 生成的试题除了第六题，题干长度均符合要求，文心大模型生成的试题频繁出现超过 11 词的题干，不符合要求。从选项长度看，真题选项一般不超过 5 词，ChatGPT 和文心大模型生成的选项长度基本符合要求。

从是否进行同义替换上看，真题或是题干、或是选项，基本会对原文表达进行了同义替换，以确保学生真正理解才能做对，而不是简单地听到词就能做对，ChatGPT 和文心大模型生成的试题在这从选项长度和类型是否一致上看，真题的选项类型均为一致，长度最多相差 1 词，ChatGPT 和文心大模型生成的试题选项类型均为一致，长度基本上一致，但都出现过某个选项长度特别突兀，而长度过长的选项均为正确选项，这增加了学生蒙对的几率，不符合要求。

Table 5. Analysis of item writing principles
表 5. 命题原则分析

题号	题干和选项是否简短	题干和选项是否对原文同义替换	选项长度和类型是否一致	干扰项是否使用文中信息命制
真题 1	是	否	是	否
真题 2	是	是	是	是
真题 3	是	是	是	是
真题 4	是	否	是	是
真题 5	是	是	是	是
真题 6	是	是	是	是
真题 7	是	是	是	是
真题 8	是	是	是	是
真题 9	是	是	是	是
真题 10	是	是	是	是
ChatGPT1	是	否	是	否
ChatGPT2	是	是	是	否
ChatGPT3	是	是	是	否
ChatGPT4	是	否	是	否
ChatGPT5	是	否	是	否
ChatGPT6	否	是	否	是
ChatGPT7	是	是	是	是
ChatGPT8	是	否	是	是
ChatGPT9	是	是	是	是
ChatGPT10	是	是	是	否
文心大模型 1	是	否	否	否
文心大模型 2	是	是	是	是
文心大模型 3	否	是	是	否
文心大模型 4	否	否	否	否
文心大模型 5	是	是	是	否
文心大模型 6	否	是	是	是
文心大模型 7	否	否	是	是
文心大模型 8	是	否	是	否
文心大模型 9	是	是	是	是
文心大模型 10	否	是	是	否

从干扰项是否使用文中信息命制上看,真题的两个干扰项中一般至少有一个是来自原文,以增强干扰效果,但尽管在提示词中明确要求,ChatGPT 和文心大模型生成试题的干扰项虽然与正确选项类型一致,但却与文本无关,一部分原因是生成文本中信息不足以命制干扰项,但即使文本中信息充足,ChatGPT 和文心大模型都倾向于不使用文中信息命制干扰项。

4.2.2. 考查的听力微技能

从考查的听力微技能上看,2023 年杭州市中考英语听力真题中,14 题考查理解细节信息的能力,1 题考查推断人物关系的能力,虽然没有出现主旨大意题和意图态度题,但这些细节理解题均符合听力单选题命题原则,试题质量高,干扰项干扰效果强,需要学生真正理解才能做对。ChatGPT 和文心大模型都能根据提示词要求,生成考查指定听力微技能的试题。在生成细节理解题方面,二者生成的干扰项合理,但并未使用文中信息,使干扰效果打折扣。在生成试题中,第 6 题为推理判断题,ChatGPT 生成的试题 6 虽然选项长度不一致,但干扰项基于文中信息,有一定难度,文心大模型生成的试题 6 正确选项是文中明确出现的信息,因此更像细节理解题而非推理判断题。第 9 题为主旨大意题,ChatGPT 和文心大模型生成的试题 9 都基本符合要求,干扰项以偏概全,有一定干扰效果。第 10 题为意图态度题,ChatGPT 和文心大模型生成的试题 10 正确选项均为褒义词,干扰项均为贬义词,因此容易排除,难度较低。

此外,文心大模型生成的试题 7 本身存在谬误,题干问作者在哪编故事并读给别人听,正确答案是在公交车上,但文中只说了作者在公交车上读故事,编故事是在家中。二者在选项中都偶尔会出现初中阶段不常见的词或超纲词,如 server, buddies, acquaintance 等,在试题投入使用前需要手动修改或重新生成。

5. 结语

总而言之,在文本生成上,ChatGPT 和文心大模型差距较小,二者都能生成话题、词数、难度与样本文本基本一致的文本。相比 ChatGPT,文心大模型生成的文本对样本文本改动较小,仅替换部分表达或扩写。此外,文心大模型容易生成词数远超过样本文本的文本,需令其重新生成。在试题生成上,ChatGPT 的表现好于文心大模型,除了二者生成试题的干扰项都倾向于不使用文中信息命制以外,文心大模型出现了更多问题,如题干过长,试题本身存在谬误等,但二者生成的试题总体质量都难以媲美真题,部分原因是本文仅使用两种人工智能第一次生成的试题进行分析,未进行反复修改。本研究的不足之处在于仅对生成文本和试题的质量进行主观分析判断,若要进一步取得关于生成式人工智能命题能力的证据,应对试题进行试测以及心理测量分析。最后,笔者认为,鉴于寻找合适听力文本的困难,英语教师在为学生编写低利害的练习题时,使用生成式人工智能来生成听力文本是一个不错的选择,但要生成高质量的试题,则需反复修改,以及需要更多人类智慧的参与。人工智能技术还在不断革新,并带来越来越大的机遇,中国应跟随时代的浪潮,继续大力发展人工智能,以达到变革和改进学校教育的目的。

参考文献

- [1] 焦建利. ChatGPT 助推学校教育数字化转型——人工智能时代学什么与怎么教[J]. 中国远程教育, 2023, 43(4): 16-23.
- [2] 平原. 生成式人工智能在高中生物学考试命题中的应用测试和问题建议[J]. 生物学教学, 2024, 49(12): 54-56.
- [3] 卢宇, 余京蕾, 陈鹏鹤, 等. 生成式人工智能的教育应用与展望——以 ChatGPT 系统为例[J]. 中国远程教育, 2023, 43(4): 24-31, 51.
- [4] 黄荣怀. 人工智能正加速教育变革: 现实挑战与应对举措[J]. 中国教育月刊, 2023(6): 26-33.
- [5] Sayin, A. and Gierl, M. (2024) Using OpenAI GPT to Generate Reading Comprehension Items. *Educational Measurement: Issues and Practice*, 43, 5-18. <https://doi.org/10.1111/emip.12590>

-
- [6] Kyung-Mi, O. (2024) A Comparative Study of AI-Human-Made and Human-Made Test Forms for a University TESOL Theory Course. *Language Testing in Asia*, **14**, Article No. 19. <https://doi.org/10.1186/s40468-024-00291-3>
- [7] Aryadoust, V., Zakaria, A. and Jia, Y. (2024) Investigating the Affordances of OpenAI's Large Language Model in Developing Listening Assessments. *Computers and Education: Artificial Intelligence*, **6**, Article 100204. <https://doi.org/10.1016/j.caeai.2024.100204>
- [8] Lin, Z. and Chen, H. (2024) Investigating the Capability of ChatGPT for Generating Multiple-Choice Reading Comprehension Items. *System*, **123**, Article 103344. <https://doi.org/10.1016/j.system.2024.103344>
- [9] Shin, D. and Lee, J.H. (2023) Can ChatGPT Make Reading Comprehension Testing Items on Par with Human Experts? *Language Learning & Technology*, **27**, 27-40.
- [10] 杨志明, 徐庆树, 祁长生. ChatGPT 命题潜力的实证研究[J]. 教育测量与评价, 2023(4): 3-13.
- [11] 梅凌宇, 吴小杰. ChatGPT 命制高中数学试题初探[J]. 中学数学月刊, 2024(6): 57-61.
- [12] 覃朝姣. 高中英语听力试题命题“三部曲” [J]. 教学考试, 2022(12): 69-72.