

【评价研究】

ChatGPT 命题潜力的实证研究

杨志明 徐庆树 祁长生

【摘要】ChatGPT 是一种功能十分强大的自然语言处理工具,若能利用它协助命题,则对考试工作和作业设计有重大意义。根据现代教育测量理论的命题要求,探讨了 ChatGPT 的命题潜力,发现在命题素材完备、任务指令明确且具体等条件下,用户通过连续追问等方式,可以诱导 ChatGPT 撰写出质量不错的试题,或提供一些富有创意的命题情境。不过,ChatGPT 撰写的客观题存在着没有答案或多个答案并存等风险,其撰写的主观题难度往往难以控制。因此,ChatGPT 撰写的题目必须通过高水平专家团队的系统审查,最好能对 ChatGPT 撰写的试题做先导测试,以确保题目的质量和难度要求。这些发现对作业设计和低利害考试的命题工作具有较大价值。

【关键词】ChatGPT;考试命题;《中国英语能力等级量表》;高考评价体系

一、ChatGPT 概述

1. ChatGPT 的产生和发展

ChatGPT 是 OpenAI 公司推出的一种自然语言处理(natural language processing, NLP)模型。其中,Chat 是对话的意思,GPT 是 generative pre-trained transformer(生成式预训练变换模型)的缩写。其工作机理是通过预先学习(pre-trained)大量现成文本和对话集合,对用户输入的文本通过生成(generative)方式做出各种智能回答(chat)。其智能水平也随着 ChatGPT 的升级换代而不断提高。

根据 OpenAI 公司发布的信息,ChatGPT 的升级主要体现在其所用模型的参数量和用于预训练的数据量的增长方面。其中,2019 年 2 月升级的 GPT-2 的参数量为 15 亿,预训练数据量为 40GB。2020 年 5 月升级的 GPT-3 的参数量为 1750 亿,预训练数据量为 45TB。2021 年升级的 GPT-3.5(参数量和预训练数据未知)可以回答任何文本问题,并完成多轮对话。2023 年 3 月升级的 GPT-4(参数量和预训练数据未知),在图像、音频等处理方面具有更好的表现,其智能水平超过了 90% 的美国高考 SAT-I 考生。

2. ChatGPT 的核心技术与功能

ChatGPT 的核心技术^[1]主要有 5 项。

(1) transformer(编码与解码转换)技术,一种基于自注意力机制的深度神经网络。这种网络包括编

码器(encoder)和解码器(decoder)两个部分,其中编码器包括自注意力子层(self-attention)和全连接前馈神经网络子层。编码器子层的作用是计算输入序列中不同位置之间的依赖关系结构并进行特征表示,对新生成的特征表示进行处理,并生成最终的特征向量。解码器的基本结构与编码器类似,但针对编码器的输出,增加了新的多头注意力层(multi-head attention),加入了参数归一化设定(add & norm),以防止解码过程中后继位置信息泄露(residual connection)。

(2) 基于 transformer 的网络结构。GPT 的信息转换器由很多层编码器和解码器组成,其最终构成一个十分庞大的神经网络。该网络运用自回归方式,实现了通过给定上文预测下文,或通过给定下文推测上文的函数,并解决了不同格式文本语料之间的转换问题。

(3) RLHF 技术。即基于人类反馈的强化学习(reinforcement learning from human feedback, RLHF)。这种强化表现为 3 种方式:其一,监督式微调,即利用符合人类预期的少量标注数据对预训练模型参数进行调整,初步优化文本生成模型;其二,构建奖励模型,目标是通过监督式微调生成的多个结果进行人工排序标记,训练奖励函数模型,用于强化学习模型输出结果的自动化评价;其三,使用近端策略优

化算法(proximal policy optimization, PPO),即结合奖励模型对文本生成模型的结果进行自动评估,并采用强化方式对文本生成模型进行优化,使其最终具备生成符合人类预期的文本的能力。

(4)指令微调技术(instruction tuning)。即通过将预设的指令描述与生成文本相结合,构建训练数据,从而微调文本生成模型的参数。其核心思想是将自然语言处理相关任务转化为基于指令描述的文本生成问题。GPT可以基于指令描述,促使模型理解指令任务,进而生成预期文本。

(5)思维链技术(chains of thought)。即有针对性地设计模型的指令输入,促使模型将单步骤推理任务拆解为包含多个中间步骤的任务。其中,每一个中间步骤由一个相对简单的指令输入作为引导,从而构成一个有逻辑的多步骤过程。这种技术可以引导文本类内容的生成,辅助模型生成和解决复杂逻辑推理任务。

与过去的AI系统相比,ChatGPT的最大特点是使用了生成式而不是基于规则的连续提问对答方式,其主要功能表现为以下4个方面。(1)创造性地生成新内容:可以根据文本中的上下文信息,产生具有创意的答案。(2)捕捉用户的意图:可以通过多轮对话捕捉用户的意图,不断修改自己的答案。(3)执行序列任务:可以通过多轮对话方式,执行用户提出的序列任务。(4)编写程序:可以根据需要编写计算机程序。^[1]

3. ChatGPT 对教育的影响

在教育领域,ChatGPT对教、学、评3个方面都有着重大影响。在教学方面,教师可以借助ChatGPT的创意文本生成功能、RLHF技术和可执行序列任务的功能,通过不断提问、缩小问题范围等方式,让ChatGPT生成教学素材、设计教学方案、模拟教学过程等。在学习方面,学生可以利用ChatGPT的互动式人机对话功能,搜寻知识、解答疑点、概括要点和探究未知等。在教学评价方面,教师可以利用ChatGPT编写试题、设计试卷、编写作业、诊断学情、撰写报告等。许多学者认为,ChatGPT可以促进教学模式转变,适应教育数字化、数字教育的转型,从传统“师—生”二元结构转向“师—生—机”三元结构;可以加速现行考核体系转变,逐步从知识测评转向“知识+素养”测评;可以推进现有智能教育应用生态的转变,从孤立、专用的教育应用转变为连接、共享、智

能的教育平台;可以促进教育理念的转变,实现跨班级、跨学科、跨学校、跨时空的个性化学习方式等。^[2]

不过,也有不少学者十分担忧ChatGPT可能引发教育诚信危机等问题。^[3]比如,学生可能用它代写作业或论文,研究人员还可能用它代写研究报告或论文,这就会干扰正常教学,突破学术伦理底线等。同时,由于ChatGPT生成的文本格式非常规范,用户不太容易察觉其可能存在的错误或误导。笔者在探究ChatGPT功能的过程中,就遇到过它编造虚假参考文献,回答问题模式化、套路化等方面的问题。这就要求教师在传授知识的同时,还要着力加强学生分析性思维和批判性思维能力的训练,培养学生识别信息真伪的习惯和能力。此外,ChatGPT还可能导致新的隐私保护问题,引发公众对个人隐私、数据安全及伦理问题的担忧。

不过,上述研究大都属于理性分析,实证研究并不太多。为了用好ChatGPT,克服其缺点,发挥它在考试行业的价值,很有必要通过实证的方式系统考察其命题水平。

二、基于现代教育测量理论的 ChatGPT 命题设计

利用ChatGPT命题,必须先给它设计一系列工作指令。按照现代教育测量学的要求,要命制优良试题,就要预先设定考试蓝图、明确命题的具体要求、提供完整的命题素材等,这些要求都需要反映在给ChatGPT的指令当中。

1. 考试蓝图要求

在设计英语试题或作业时,首要任务是要设计考试蓝图,界定试题或作业所要覆盖的必备知识、英语能力和学科素养。根据《中国英语能力等级量表》(简称CSE),英语知识主要包括语法知识、篇章知识、功能知识和社会语言知识4个模块,英语能力包括语言理解能力(听力和阅读)、语言表达能力(口语和写作)、语用能力、翻译能力和语言使用策略5个维度^[4];根据课程标准,英语学科核心素养包括语言能力、文化意识、思维品质和学习能力4个方面^[5-6]。下页表1^[7]是基于CSE和学科核心素养设计的英语考试蓝图。

根据表1,本研究要求ChatGPT分别创设一些能够考查英语语法知识、篇章知识、功能知识和社会语言知识的题目,同时要求这些题目能覆盖课程标准所要求的语言能力、文化意识、思维品质和学习能力4种学科核心素养。

表 1 基于 CSE 和学科核心素养的英语考试蓝图设计样例

统计量		语言能力					文化 意识	思维 品质	学习 能力	总分
		语言理解能力	语言表达能力	语用能力	翻译能力	语言使用策略				
英语知识	语法知识	7	7	2	0	2	4	4	5	31
	篇章知识	5	3	2	0	2	4	4	4	24
	功能知识	3	3	2	0	2	4	4	4	22
	社会语言知识	3	4	2	0	2	4	4	4	23
合计		18	17	8	0	8	16	16	17	100

注:表格中的数字可以是分数比例或题目数量。

2. 考查要求

高考评价体系提出了基础性、综合性、应用性和创新性的考查要求。^[8]因此,命题人员在给 ChatGPT 发出指令时,可以明确要求它分别命制一些基础性的、综合性的、应用性的、创新性的试题。^[9]

基础性命题主要考查学生对英语学科主干知识的掌握程度,包括对语法知识、篇章知识、功能知识和社会语言知识的掌握程度;综合性命题考查的是学生对所学知识的融会贯通程度,以及综合运用听、说、读、写等多种英语能力的水平;应用性命题主要体现在真实或模拟的语言情境中运用所学英语知识的水平,包括学生使用英语从事日常交际、学习和思辨活动,达到交际目的和解决实际问题的能力;创新性命题强调创新意识和创新思维,要求合理呈现试题情境,设置新颖的试题呈现方式和设问方式,考查学生完成开放性或探究性任务的能力,鼓励学生创造性地思考问题和解决问题,包括引导考生积极参与跨文化交流、学习和思辨活动,积累中外文化知识,掌握多种思维方式,在跨文化的互动与融合中形成一定的创新意识和能力。

3. 命题素材要求

为了方便,本次命题的素材选择了《新概念英语》第2册第1课的内容,设计了3种水平的素材信息。水平1:只要求 ChatGPT 以新概念英语第2册第1课的内容为命题素材;水平2:给 ChatGPT 提供课文全文;水平3:给 ChatGPT 提供课文全文以及相关词汇的解读。

三、基于现代教育测量理论的 ChatGPT 命题表现

参照 CSE 和高考评价体系,利用 ChatGPT 命题的英语试题内容主要包括语法知识、篇章知识、功能知识和社会语言知识,考试命题主要从基础性、综合性、应用性或创新性方面进行设计。以下分别针对

客观题(4 选 1 题型)和主观题的命题设计及其效果进行讨论。

1. ChatGPT 命制客观题的表现

(1)若用户给 ChatGPT 的指令比较抽象,且所用命题素材信息为水平 1,则 ChatGPT 的命题表现不够理想。在这种命题情境下,它撰写的题目不仅难度不够,而且可能存在错误,并与用户给定的命题素材基本无关。比如,仅仅要求 ChatGPT 针对《新概念英语》第2册第1课的素材,从语法知识、篇章知识、功能知识和社会语言知识 4 个方面分别命制 2 题,它给出了 8 题(这里只选取 4 题呈现,题 1~题 4)。

题 1(语法题/基础性命题): Which sentence is grammatically correct?

- A. I plays football every day.
- B. She play football every day.
- C. He plays football every day.
- D. They play football every day.

题 2(语法题/基础性命题): Choose the correct form of the verb to complete the sentence: We _____ our homework yesterday.

- A. do. B. did. C. does. D. done.

题 3(篇章题/综合性命题): What is the main topic of the passage?

- A. Sports. B. Travel. C. Food. D. Education.

题 4(篇章题/综合性命题): What is the author's purpose in writing this passage?

- A. To entertain readers.
- B. To persuade readers.
- C. To inform readers.
- D. To criticize something.

【题目点评】

题 1 和题 2 考查的是语法知识,分别考查动词

的形态与时态如何与主语保持一致的问题,但第1题的选项中有两个答案(C和D),第2题答案为B。这两道题目质量不太好,而且与命题素材无关。

题3和题4考查的是篇章知识,分别考查阅读材料的“主旨”和作者的“意图”,但题3没有正确答案,题4答案为D。这种题目似乎是一种套路。

(2)若用户给 ChatGPT 的指令比较明确和具体,所用命题素材信息为水平2,则 ChatGPT 的命题水平会有所提升,但质量仍然不尽如人意,而且题目内容与用户给定的命题素材无关。例如,给 ChatGPT 提供《新概念英语》第2册第1课的原文素材,并要求它从语法知识、篇章知识、功能知识和社会语言知识4个方面分别命制2题,它给出了8题(这里只选取4题呈现:题5~题8)。

题5(语法题/基础性命题): Choose the correct form of the verb to complete the sentence: I _____ a new book yesterday.

- A. buy B. buys C. bought D. buying

题6(语法题/基础性命题): Which sentence is grammatically correct?

- A. She don't like pizza.
B. He doesn't likes pizza.
C. They doesn't like pizza.
D. We don't like pizza.

题7(篇章题/综合性命题): What is the main topic of the passage?

- A. Daily routines.
B. Sports activities.
C. Famous landmarks.
D. Food preferences.

题8(篇章题/综合性命题): What is the author's purpose in writing this passage?

- A. To entertain readers.
B. To persuade readers.
C. To inform readers.
D. To describe a personal experience.

【题目点评】

题5和题6考查的是语法知识,分别考查动词的时态和助动词用法,答案分别为C和D。这两道题目很简单,而且与命题素材无关。

题7和题8考查的是篇章知识,分别考查阅读材料的“主旨”和作者“意图”,但题11仍然没有正确答

案,题12的答案勉强为D。两题都与命题素材无关。

(3)若用户给 ChatGPT 的指令非常具体,所用命题素材信息为水平为3,则 ChatGPT 的命题表现非常不错。它撰写的题目不仅与指定的命题素材密切相关,而且能撰写出较好的创新性试题。例如,当笔者不仅给 ChatGPT 提供《新概念英语》第2册第1课的原文,而且提供了原文的相关注释,再要求它从语法知识、篇章知识、功能知识和社会语言知识4个方面分别命制2题时,它给出了8题(这里只选取4题呈现:题9~题12)。

题9(功能题/创新性命题): Based on the passage, why did the narrator say, "I can't hear a word!"?

- A. They genuinely couldn't hear the actors.
B. They wanted to start a conversation with the young man.
C. They wanted to express their anger to the young man and woman.
D. They were trying to be polite to the actors.

题10(功能题/创新性命题): What was the young man's intention when he said, "This is a private conversation!"?

- A. He wanted to apologize to the narrator.
B. He wanted the narrator to join their conversation.
C. He wanted the narrator to be quiet.
D. He wanted to explain his behavior to the narrator.

题11(社会语言题/创新性命题): In the given context, why did the young man's response seem rude?

- A. He used impolite language.
B. He interrupted the narrator.
C. He ignored the narrator's request.
D. He raised his voice.

题12(社会语言题/创新性命题): Which word could be used to describe the young man's behavior?

- A. Formal. B. Polite. C. Respectful. D. Rudely.

【题目点评】

题9和题10考查的是功能知识,分别考查考生对原文作者所说的“一句话也听不清”以及年轻男子回答的“这是私密谈话”所暗示的意思,答案都为C。题目都与命题素材密切相关,并且很好地考查了特

定语境下话中有话的真正意思,但题9中的 They 需要修改成 He, their 要改为 his, 命题质量一般。

题11和题12考查的是社会语言知识,分别询问年轻男子反应无礼的原因以及如何形容这个年轻男子行为的问题,二者答案分别为C和D。题目与命题素材密切相关,考生只有具备一定的社会语言知识才能作答正确,表现出了良好的命题水平。

2. ChatGPT 命制主观题的表现

若给 ChatGPT 发出十分具体的工作指令,同时提供丰富的命题素材,并要求它撰写综合性和创新性试题,则可以得到相当不错的主观性试题。ChatGPT 所创设的问题情境很有启发意义。以下是利用 ChatGPT 所具有的连续对话功能,采用不断提问、聚焦话题等方式获得主观性试题的过程。

首先,给 ChatGPT 做出一定的角色设定。综合性题目与创新性题目的定义抽象程度较高,需要首先对 ChatGPT 做专家设定。“You are well versed in writing test items for assessing the students' competence in terms of English as a second language. Now I will give you several passages. You are required to generate some innovative items and some items that requires a composite of competences including syntactical knowledge and pragmatic knowledge based on the materials provided.”(“你非常善于撰写英语二学习得方面的题目,现在我会提供给你一些文章,请你利用给定素材生成一些创新性的题目或者综合性的题目。”)。

其次,为 ChatGPT 提供具体的命题素材。本研究为 ChatGPT 提供了《新概念英语》第二册第1课的原文及其注释。于是,ChatGPT 首先给出了一些以考查知识点为主的客观性试题(提问指令没有强调撰写主观题的要求)。

例1(创新性命题):What does the word “bear” mean in the context of the passage?

- A. Large furry mammal.
- B. Tolerate or endure.
- C. Expressing anger.
- D. Direction indicator.

例2(综合性命题):Identify the error in the following sentence: “I did not enjoyed it.”

- A. The verb should be “enjoyed”.
- B. The word “it” is unnecessary.

C. The sentence is correct.

D. The verb should be “enjoy”.

其中,例1创新性试题主要考查词汇内容,例2综合性试题仅仅考查语法知识。

随后,通过进一步更新指令,对 ChatGPT 不断提出研发创新性和综合性题目的要求,获得了越来越复杂的主观性试题。

例3(创新性主观题):Write an evocative and detailed review of the play from the perspective of the narrator. Use expressive language to convey the emotions experienced during the performance and how the disruptive incident impacted their overall impression of the play.

例4(综合性主观题):Rearrange the words to form a grammatically correct sentence based on the passage: “loudly/talking/a/young man/woman /The/and/were/behind me.”

例5(创新性主观题):Imagine you were the narrator in this situation. Instead of confronting the young man and woman directly, suggest an alternative approach you could take to address the issue while maintaining a polite atmosphere in the theater.

例6(综合性主观题):Present a scenario where the theater manager must decide whether to prioritize artistic freedom or audience comfort. Formulate a well-structured argument that considers the ethical implications of potential decisions and the long-term effects on the theater's reputation and patronage.

由以上例题可知,通过连续追问,ChatGPT 给出的题目越来越复杂,难度也越来越大。比如,在创新性命题指令下,ChatGPT 要求答题者从原文作者的角度对戏剧演出做出评论,并使用表达性较好的语言词汇对表演过程中原文作者的情感体验进行描述;在综合性命题指令下,ChatGPT 要求答题者从剧院经理的角度考虑问题,对艺术自由和观众的舒适度的关系进行讨论,并顾及决策的伦理、对剧院名声和获得赞助的长期影响等因素。

值得一提的是,例4设计得不够好,其中的“The”最好能修改成“A”,其正确的答案可以是“A young man and a woman behind me were talking loudly.”,若直接使用 ChatGPT 给出的题目,则答案可勉强写成“The man and a young woman behind me were talking loudly.”(也不太合适)。

从上述命题表现来看,ChatGPT 具备了撰写创新性和综合性主观题的能力,它可以在戏剧评论、句子构建、场景设计、观点讨论、社交礼仪等方面设计题目,加大了语言应用、思维技能和综合素养方面的考查力度,与传统的过度关注语法知识和词汇知识的考查方式明显不同。

不过,ChatGPT 设计的题目也存在一些问题,例如,这些题目大都涉及短文写作任务,对于英语初级水平的学习者而言明显偏难。事实上,这些题目不仅仅考查了学生的英文水平,也对他们的人文素养和思维技能方面的水平提出了很高的要求。要回答好这些问题,学生不仅需要涉猎各方面知识,而且需要具备较高的分析性思维、批判性思维和创造性思维水平。另外,ChatGPT 目前在运用多媒体手段方面的条件还不足,无法生成基于图片或音频、视频的题目。

四、启示与建议

根据上述实证分析,可以得到如下启示和建议。

首先,在命题素材完整、指令要求科学且具体的情况下,ChatGPT 可以撰写出一些高水平的基础性、综合性、应用性、创新性的客观性试题。退一步讲,ChatGPT 至少能为撰写高质量试题提供富有创意的命题情境。

其次,利用 ChatGPT 所具有的生成式多轮对话功能,可以让它撰写出一些难度很大的综合性和创新性主观性试题。

再次,ChatGPT 的当前版本具有浓厚的话语接龙特征,它所命制的题目很可能出现科学性错误,存在着没有答案或多个答案并存等风险。因此,ChatGPT 撰写的所有题目,都需要由高水平专家团队,通过人工审查等方式,对其内容的科学性和恰当性等进行系统检查。

最后,ChatGPT 命制的试题难度不太好把握,容易出现非常容易或非常难的试题。因此,ChatGPT 所撰写的题目最好能通过先导测试,或采用小范围内高度保密条件下的试测手段,加大心理计量学分析(psychometric analysis)力度,确保试卷上所有题目不存在科学性问题的,较好地反映考试蓝图所设定的要求,以保证考试的内容效度。

总之,ChatGPT 在考试命题方面潜力很大。只要命题要求明确、合理,命题素材完整,它就能撰写出质量不错的客观性试题和高难度的主观性试题,至少可以为命题教师提供富有创造性的问题情境,

打开命题思路,为提高命题质量提供有力支持。美中不足的是,ChatGPT 所撰写的题目可能存在科学性错误,其难易程度不太好把握,因此,ChatGPT 所撰写的考试题目,还需要由一个高水平专家团队来把关,最好可以对 ChatGPT 所撰写的试题在使用之前进行先导测试,以确保试题质量。当然,若仅仅把它撰写的题目用于日常作业或低利害的诊断性考试,ChatGPT 则是大有可为的。

参考文献:

- [1]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.
- [2]杨宗凯,王俊,吴砥,等.ChatGPT/生成式人工智能对教育的影响探析及应对策略[J].华东师范大学学报(教育科学版),2023,41(7):26-35.
- [3]黄荣怀.人工智能正加速教育变革:现实挑战与应对举措[J].中国教育旬刊,2023,(6):26-33.
- [4]教育部考试中心.中国英语能力等级量表[M].北京:高等教育出版社,上海:上海外语教育出版社,2018:1-5.
- [5]教育部.义务教育英语课程标准(2022年版)[S].北京:北京师范大学出版社,2022:4-5.
- [6]教育部.普通高中英语课程标准(2017年版2020年修订)[S].北京:人民教育出版社,2020:4-5.
- [7]杨志明,顾兰,夏胜俊.矩阵式过程评价量表设计初探:以英语学科为例[J].中国考试,2023,(8):28-36.
- [8]教育部考试中心.中国高考评价体系[M].北京:人民教育出版社,2019:28-32.
- [9]陈康,吴泓霖,李新煜,等.基于高考评价体系的英语科考试内容改革实施路径[J].中国考试,2019,(12):33-37.

【作者简介】杨志明,湖南师范大学测评研究中心主任,外国语学院教授,博士生导师,香港中文大学博士(长沙 410081);徐庆树,湖南师范大学外国语学院博士生;祁长生,深圳市卓帆技术有限公司董事长。

【原文出处】摘自《教育测量与评价》(长沙),2023.4.3~13

【基金项目】本文为中国教育发展战略学会教育考试专业委员会2021-2022年度教育考试专项课题“智能化自适应学业水平考试系统的研发”(jyks2022021)阶段性成果。