Advances in Education 教育进展, 2023, 13(11), 8213-8224
Published Online November 2023 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2023.13111273

Hans汉斯

大语言模型背景下提示词工程赋能英语口语学习研究

郭子浩，孙由之，张梦林，王欣然，陈雨洁

中国矿业大学(北京)文法学院，北京

收稿日期：2023年9月30日；录用日期：2023年10月28日；发布日期：2023年11月2日

摘　要

本研究探讨了运用提示词工程和大语言模型来适应各种英语口语教育场景，研究了为模拟雅思口语考试、K-12口语考试和学前英语教学等场景而定制的提示词的设计和应用。本研究使用了定量指标，如BLEU和ROUGE分数，以及流利性、相关性、完整性、多样性和连贯性等指标，评估生成的回复的质量。此外，本研究还运用了综合评分公式来评估回复的质量，并说明了每个指标的重要性。本研究强调了提示词工程在为英语教育提供适应性解决方案方面的潜力。它还强调了需要进一步改进大语言模型的能力，以提高其在这些场景中的性能。本研究为提示词工程和大语言模型的应用提供了宝贵的见解，为英语口语教育及相关领域的发展做出了贡献。

关键词

英语口语学习，语言大模型，提示词工程，人工智能

Empowering Spoken English Learning with Prompt Engineering against the Background of Large Language Models

Zihao Guo, Youzhi Sun, Menglin Zhang, Xinran Wang, Yujie Chen

School of Law and Humanities, China University of Mining and Technology (Beijing), Beijing

Received: Sep. 30th, 2023; accepted: Oct. 28th, 2023; published: Nov. 2nd, 2023

Abstract

This research explores the utilization of prompt engineering and large language models to adapt

文章引用: 郭子浩, 孙由之, 张梦林, 王欣然, 陈雨洁. 大语言模型背景下提示词工程赋能英语口语学习研究[J]. 教育进展, 2023, 13(11): 8213-8224. DOI: 10.12677/ae.2023.13111273

to various English oral education scenarios. It investigates the design and application of prompts tailored to simulate scenarios such as IELTS speaking tests, K-12 oral exams, and preschool English teaching. The study assesses the quality of generated responses using quantitative metrics, including BLEU and ROUGE scores, fluency, relevance, completeness, diversity, and coherence. Furthermore, it introduces a comprehensive scoring formula to evaluate response quality, accounting for the significance of each metric. Despite some limitations, the research highlights the potential of prompt engineering in providing adaptable solutions for English language education. It also underscores the need for further improvements in LLMs' capabilities to enhance their performance in these scenarios. This study contributes to the advancement of English oral education and related fields by providing valuable insights into the application of prompt engineering and large language models.

Keywords

Spoken English Learning, Large Language Models, Prompt Engineering, Artificial Intelligence

1. 引言

1.1. 背景与动机

在中国的外语口语学习领域，历经三个重要阶段。首阶段，被赋予“1.0 版本”的标签，见证了口耳相传与录音机为主要教学手段的原始时期。次阶段，随着电子工艺的飞速进步，学习机、复读机以及点读机等技术应运而生，有效地丰富了语料库资源并提升了学习的互动性。然而，当人工智能小型模型技术引入后，外语学习进入了第三个阶段，呈现出更高效和便捷的趋势。这个阶段不仅涌现了备受瞩目的产品如流利说英语、多邻国等，也为用户提供了优质的学习服务和体验。尽管如此，在口语陪练领域，小型模型的规则限制仍然制约着用户对话的自由度[1]。

然而，随着技术的蓬勃发展，大语言模型(LLM)的崭露头角，其中 CHATGPT 更是名声在外。这类 LLM 不仅为各行业带来了新的可能性，尤其在语言教育领域产生了深远的影响[2]。然而，在当前国内市场上，仍然鲜见能够根据用户需求量身定制口语训练方案，并为用户提供智能化口语陪练服务的产品或软件[3]。更进一步，现有产品往往难以在技术和产品层面提供足够灵活、足够智能、足够人性化的体验，以迎合用户多样化的学习需求。然而，正是 LLM 技术的不断演进，为解决这一系列问题提供了难得的机遇和技术支持。通过对历史背景的深入分析，我们能够更好地理解外语口语学习领域的发展脉络，并在这个背景下，更准确地勾勒出本研究的动机和目标。本研究旨在充分利用 Prompt 工程的概念和 LLM 技术的优势，探索如何在英语口语教育中创新应用，突破小型模型的限制，为用户提供更高水平的口语学习体验[4]。

1.2. 目标与意义

1.2.1. 研究目标

本研究的核心目标在于探索基于大语言模型(LLM)和提示词工程的英语口语教育研究领域，填补这一领域在国内研究中的空白。以提示词工程为研究核心，我们旨在深入研究多种口语学习场景，对口语

题库和教学策略进行深入分析，并将英语教育与大模型技术相融合，为英语口语教育领域带来新的创新性贡献。

1.2.2. 研究意义

本研究具有重要的理论和实践意义，具体体现在以下几个方面：

一是填补研究空白与创新性贡献。当前，国内关于基于 LLM 和提示词工程的英语口语教育研究仍处于相对空白的状态。通过本研究的深入探索，我们将弥补这一领域的研究空白，为该领域的理论体系构建和实践应用提供有力的支持。通过对多种口语学习场景的调研与分析，我们能够挖掘出更加切合实际的教学策略，从而为英语口语学习与教学提供更为精准和有效的指导。

二是推动英语口语教育的技术创新。提示词工程作为本研究的核心概念，将大语言模型与教育教学相结合，为英语口语教育的技术创新提供了新的路径。通过对提示词设计、口语题库构建等方面的研究，我们将探索如何更好地应用 LLM 技术，以提升口语学习的互动性、个性化和效率。这对于促进英语口语教育的现代化转型具有积极的推动作用。

三是实践应用价值与教育改进。本研究不仅停留在理论层面，更着眼于实际应用价值。我们将通过对基于 LLM 和提示词工程的交互式英语口语学习 APP 的构建，为广大学习者提供一个创新的学习平台。通过该平台，学习者可以获得个性化的口语学习方案，并在智能化陪练的引导下，提升自己的口语表达能力。这将有望为英语教育的质量提升和学习效果改进作出实际贡献。通过本研究的深入探索，我们将在理论和实践层面为英语口语教育领域带来新的视角和方法，促进该领域的持续发展和创新。同时，研究成果还将为教育实践者、研究者和决策者提供有益的参考，以推动英语教育的进步和优化。

1.3. 研究方法概述

提示词工程实验方法

提示词组成原理明确：本研究首先对提示词工程的组成原理进行了详细的解析，涵盖了 background、role、context、case、instruction、limit、Associated data 等关键要素。通过深入理解每个要素在形成有效提示词时的作用，确保所构建的提示词能够在不同情境下引导出准确的口语表达。

应用搭建与实验：基于百度飞桨开放平台 AISTUDIO，我们搭建了一个可供实验的应用平台，用以探索提示词工程的效用。在平台上，我们进行了一系列实验，通过不断的参数调整和优化，达到了实验效果的最佳状态。这一步骤旨在验证提示词工程的可行性，以及对于英语口语学习的有效性。

批量化复制：通过在实验中获取的关键数据和经验，我们成功地将大量英语口语对话场景进行了批量化复制。通过补齐关键字段如用户画像、题目信息等内容，我们实现了对多样化对话情境的高效生成。这一步骤为英语口语学习的实际应用提供了可行的解决方案。

2. 大语言模型(LLM)与提示词工程

2.1. LLM 的基本概念与发展

大语言模型(LLM)是当今人工智能领域的一项重要创新，其在自然语言处理任务中表现出色，成为了各领域研究和应用的热门方向。LLM 基于深度学习技术，拥有数十亿甚至上百亿的参数量，使得其在理解和生成自然语言文本方面取得了显著突破。通过大规模的预训练和微调过程，LLM 能够从大量的文本数据中学习语法、语义、情感等信息，从而实现对复杂自然语言任务的高水平处理。

LLM 在自然语言处理领域引起了巨大关注，其应用领域涵盖了文本生成、文本理解、情感分析、机器翻译等多个领域。CHATGPT 作为其中的杰出代表，通过深层的变换和自注意力机制，能够生成流畅、连贯且有逻辑性的文本，使得其在各种语言任务中都展现出令人印象深刻的性能。

2.2. 提示词工程与应用

提示词工程作为一种创新的方法，致力于利用 LLM 的强大能力，通过适当的提示词语来引导模型生成特定类型的输出。提示词工程结合了领域知识、语境设定和问题引导，以确保 LLM 能够按照期望的方式生成内容。在英语口语教育领域，提示词工程的应用为创新教学和学习方式带来了新的思路。

通过对话历史的分析和挖掘，我们可以更好地理解提示词工程的优势和应用。提示词工程能够根据用户需求，构建合适的提示词语，从而引导 LLM 生成与口语学习相关的内容。这种方式可以通过灵活调整提示词的构成原理，使得 LLM 输出更贴合不同场景的口语学习需求。正是这种灵活性和个性化定制的特点，使得提示词工程成为英语口语教育研究中的有力工具。

综上所述，大语言模型(LLM)作为一种强大的自然语言处理技术，以其深度学习的能力在多个领域展现出卓越的表现。提示词工程则进一步将 LLM 的能力延伸，为英语口语教育领域带来了全新的探索和应用。在接下来的章节中，我们将更详细地介绍如何运用提示词工程的概念和实验过程，展示如何实现英语口语学习的创新和提升[5]。

3. 适配多种英语口语场景的提示词工程实验

3.1. 提示词工程实验设计与构建

3.1.1. 场景选择与分析

在本研究中，我们针对英语口语教育领域中的不同学习场景进行了深入选择与分析，涵盖了雅思口语、K12 应试和学龄前教案设计这三个关键场景。每个场景都存在特定的痛点和需求，以下是对这三个场景的详细分析：

一是雅思口语场景。雅思口语考试作为国际认可的英语水平测试，吸引了大量备考者的关注。然而，目前市面上主要的解决方案仍然是真人对练，而这种模式存在一系列痛点：1) 高昂的费用限制了学习者的选择。针对预算有限的备考者，昂贵的真人对练费用成为一大负担，使得部分考生难以获得高质量的口语训练。2) 时间和地点限制影响学习灵活性。由于真人对练需要在特定的时间和地点与教师或陪练者进行，备考者的学习时间和地点受到较大限制，难以实现自由、灵活的学习安排。3) 交流主题和时机受限。学生只能在特定的时刻和题材上与老师进行交流，导致了缺乏在不同话题上的多样性和个性化训练，限制了备考者的学习广度和深度。

二是 K12 应试场景。近年来，国家对英语口语的重视在 K12 教育领域日益增加。具体而言，英语科目新增的口语考试在统考中占据重要地位，对学生的综合英语素养提出了更高的要求：1) 口语考试成为分数的重要组成部分。国家在英语科目中增加口语考试的权重，明确了口语在学生综合成绩中的重要地位，加强了对学生口语能力的考察。2) 强调学好口语的紧迫性。高中学生需要通过口语考试来展现自己的英语水平，而英语的实际运用对于日常交流、学术交流和社会生活等方面都有重要意义。

三是学龄前英语教育场景。学龄前儿童英语启蒙教育是家长关注的重要领域。在这个场景下，家长的需求表现为以下几个方面：1) 早期接触英语。家长希望让孩子在早期就开始接触英语，培养他们的英语语感，以便在未来的学习中更加自信和积极。2) 轻松愉快的学习氛围。家长期望孩子在学习英语的过程中能够保持兴趣，愿意在轻松愉快的氛围中进行学习，避免过于严厉和压力。3) 个性化教学。家长希望教育机构或教师能够根据孩子的兴趣和特点提供个性化的教学，激发孩子的学习兴趣和动力，确保他们在快乐学习的同时得到有效的启蒙。

通过对以上三个不同的英语口语学习场景的分析，我们深刻认识到每个场景所面临的痛点和需求，为后续的提示词工程实验过程提供了有益的指导和方向。

3.1.2. 提示词工程设计与实验准备

在进行适配多种英语口语场景的提示词工程实验前，我们对每个场景的提示词进行了详细的设计与构建。提示词的组成原理包括了 background、role、context、case、instruction、limit 和 Associated data，每个元素的具体含义和应用如表 1 所示。

Table 1. Prompt word elements and their meanings
表 1. 提示词元素及其含义

元素	具体含义
背景(Background)	背景部分描述了特定的口语学习场景，涵盖了雅思口语、K12 应试和学龄前教案设计等。这有助于为大语言模型提供关于该场景的基本背景信息。
角色(Role)	角色部分告知大语言模型需要在特定场景中扮演的角色，例如考生、学生、老师等。这有助于模型在生成回复时更好地理解自己的身份和定位。
上下文(Context)	上下文部分为模型提供了参考的背景信息，使其能够更好地理解当前对话的语境和历史。在不同场景下，上下文的内容会有所不同，有助于提供更准确的回复。
示例(Case)	示例部分提供了特定场景下的实际对话范例，用于进行 few-shot 学习。这使得模型能够从少量示例中推断出适用于特定场景的语言模式和表达方式。
任务指示(Instruction)	任务指示部分明确了模型需要完成的具体任务，例如回答问题、展开对话等。这对于引导模型生成与特定场景相关的回复至关重要。
进一步限制(Limit)	进一步限制部分对任务指示进行进一步的限制，以提高生成回复的鲁棒性。例如，限制回复的长度、强调要求回复明确、提供特定角度等。
相关数据(AssociatedData)	相关数据部分指定了与特定场景相关的数据源，这些数据可以帮助大语言模型更好地理解场景和上下文，从而生成更合适的回复。

通过以上对提示词各个组成原理的应用，我们能够设计出能够更好地适应不同英语口语学习场景的提示词。这些提示词不仅能够引导大语言模型生成与特定场景相关的内容，还能够通过 few-shot 学习提升模型的适应性，同时通过限制和关联数据的设置，提高模型输出的准确性和鲁棒性。在实验中，我们将基于这些设计原则构建不同场景下的提示词，以探究其在适配多种英语口语场景中的效果和表现。

3.2. 提示词实验内容

3.2.1. 平台选择与搭建

为了进行本次适配多种英语口语场景的提示词工程实验，我们选择了适当的实验平台和辅助平台，以保证实验的准确性和有效性。以下是我们选择和搭建的平台：

1) 实验平台：百度飞桨开放平台 AISTUDIO。我们主要选择了百度飞桨开放平台 AISTUDIO 作为我们的实验平台。AISTUDIO 提供了强大的人工智能开发环境，为我们提供了一个可以进行大规模实验和测试的环境。我们将在 AISTUDIO 上搭建实验框架，导入我们设计的提示词，进行对话交互，并收集模型生成的回复数据。

2) 大模型的对照实验平台：chathub.gg。为了进行对比实验，我们选择了 chathub.gg 作为大语言模型的对照实验平台。在 chathub.gg 上，我们将使用 ChatGPT、newbing、bard、Claud 等大模型，以进行与我们设计的提示词模型的回复进行对比和评估。

3) 大模型提示词调优平台：promptperfect。为了进一步优化设计的提示词，我们使用了 promptperfect 平台进行人工调优。promptperfect 提供了机器辅助的提示词调优环境，可以帮助我们在大语言模型的输出中进行选择和调整，以获得更加准确和适配的回复。

通过以上平台的选择和搭建，我们可以在合适的环境中进行实验、对比和调优，从而得到有关适配多种英语口语场景的提示词工程的详细实验结果和结论。这有助于验证我们设计的提示词的效果，以及与其他大语言模型的比较情况。

3.2.2. 实验内容

使用 JSON 格式，将提示词封装，用于直接发送给后端调用大模型 API 接口。tempelate 格式需要与大模型厂商对齐，本文仅作示例。

1) 雅思雅思口语场景

在雅思口语场景的实验中，我们设计了以下的提示词内容，以模拟考官与学生之间的对话交互，直接实现了用户通过大模型的对话，模拟雅思考试的过程。

```
{
"background": "You are an IELTS speaking test examiner.",
"role": "I am the student taking the IELTS speaking test.",
"sex": "Not specified",
"character": "Not specified",
"context": "You are in the context of an IELTS speaking test.",
"case": "You need to ask the following test questions to the student:",
"instruction": "Ask each question one at a time, using the exact phrasing in the parentheses. After receiving the student's response, move on to the next question.",
"limit": "Ask only one question at a time.",
"Associated_data": "N/A"
}

{
"background": "English language education and assessment",
"role": "Prompt Engineer",
"sex": "Not specified",
"character": "Professional and detail-oriented",
"context": "Language assessment criteria for spoken communication",
"case": "Language assessment and evaluation criteria",
"instruction": "Evaluate the grammar and fluency of responses",
"limit": "None specified",
"associated_data": "None specified"
}
```

2) 学龄前英语教育场景

在学龄前教案设计的实验中，我们设计了以下的提示词内容，“N/A”置空的内容用于用户自主填写的字段，提高模型返回的精准性和个性化。

{
"background": "English teaching for young learners",
"role": "Prompt Engineer",
"sex": "N/A",
"character": "N/A",
"context": "N/A",
"case": "Lesson Plan: Teaching Color Words",
"instruction": {
"teaching_goals": "Teach young learners to recognize and remember the word 'red'. Consolidate learning through games and songs.",
"teaching_preparation": [
"Prepare red objects like red apples.",
"Print or prepare relevant pictures for visual aid."
],
"teaching_steps": {
"introduction_activity": "Introduce the activity by showing a red apple and asking students to identify the color. Guide them to say 'red'. Display a card with the word 'red' and point to the letters R and E, encouraging pronunciation imitation.",
"game_activity": "Play a color matching game with red, pink, and yellow items. Students choose an item, say its color, and continue until all items are matched.",
"consolidation_activity": "Teach a color-related song, such as 'Rainbow Song'. Emphasize the color 'red' during the song. Use gestures or props to aid memorization and comprehension."
}
},
"limit": "N/A",
"associated_data": {
"song_name": "Rainbow Song",
"early_education_video": "English version of 'Rainbow Song'"
}
}

3) K12 应试场景

针对 K12 应试场景中，我们选取了单词辨析这个功能单元进行了详细实验。few-shot (少样本提示词)部分的引用，用于帮助模型更好地理解任务。

{
"background": "You are an experienced English teacher specialized in assisting students aged 6-18 in distinguishing between English words. Your role involves guiding students to differentiate between words with clear and simple explanations, adhering to the vocabulary requirements of high school students.",
"role": "English teacher for word distinctions",
"context": "You are providing guidance on distinguishing between the words 'glimpse' and 'glance' with appropriate English definitions and explanations.",

"instruction": "Your task is to explain the differences between the words 'glimpse' and 'glance' using English definitions and descriptions. Ensure that your explanations are clear, concise, and suitable for high school students.",
"Associated data": {
"word_pairs": [
{
"word": "glimpse",
"part_of_speech": "noun",
"definition": "a quick look, a brief or incomplete view, a vague indication",
"usage": "I caught a glimpse of the sunset through the trees."
},
{
"word": "glimpse",
"part_of_speech": "verb",
"definition": "to catch a glimpse of or see briefly",
"usage": "I glimpsed the famous actor as he walked by."
},
{
"word": "glance",
"part_of_speech": "noun",
"definition": "a quick look",
"usage": "She gave a quick glance at the clock."
},
{
"word": "glance",
"part_of_speech": "verb",
"definition": "to throw a glance at; take a brief look at; rebound after hitting",
"usage": "He glanced at the newspaper headlines."
}
]
}
}

简单生活场景的对话训练，我们选取了书籍对话功能进行实验。

{
"background": "I am a book enthusiast who loves engaging in English conversations with students.",
"role": "As a prompt engineer, I create responses for English language education discussions.",
"sex": "N/A",
"character": "I have a passion for books and enjoy discussing literary topics with students.",
"context": "The context revolves around discussing books and engaging students in conversations about literature.",

"case": {
"background": "I am a book enthusiast who loves engaging in English conversations with students.",
"role": "As a prompt engineer, I create responses for English language education discussions.",
"sex": "N/A",
"character": "I have a passion for books and enjoy discussing literary topics with students.",
"context": "The context revolves around discussing books and engaging students in conversations about literature.",
"instruction": {
"1": "All responses are in English using simple and understandable words, suitable for high school-level vocabulary.",
"2": "Engage in conversations around books, ask students questions, and provide related prompts to encourage further discussion.",
"3": "Ignore responses with less than 5 words; continue with new book-related questions.",
"4": "Offer evaluations of books, including story plots, character traits, political metaphors, and historical context. You can share your opinions about the books and authors, and expand on related topics.",
"5": "Introduce books with similar themes."
},
"limit": "N/A",
"associated_data": "N/A"
},
"instruction": {
"1": "All responses are in English using simple and understandable words, suitable for high school-level vocabulary.",
"2": "Engage in conversations around books, ask students questions, and provide related prompts to encourage further discussion.",
"3": "Ignore responses with less than 5 words; continue with new book-related questions.",
"4": "Offer evaluations of books, including story plots, character traits, political metaphors, and historical context. You can share your opinions about the books and authors, and expand on related topics.",
"5": "Introduce books with similar themes."
},
"limit": {
"min_words": 5
},
"associated_data": {
"books_mentioned": [
"The Three-Body Problem by Liu Cixin",
"Beijing Folding by Hao Jingfang",
"The Legend of the Condor Heroes by Jin Yong"
],
"awards_mentioned": [

"Nebula Award",
"Hugo Award"
],
"authors_mentioned": [
"Liu Cixin",
"Hao Jingfang",
"Jin Yong"
],
"themes_mentioned": [
"Sci-fi literature",
"Martial arts stories"
]
}
}

3.3. 实验评估方法

评估一个提示词的质量涉及多个因素，设计一个用于打分测试的公式和定量指标是一项复杂的任务，因为它需要综合考虑多个因素，包括生成的文本的质量、相关性、流畅性等等。以下是一些可能用于评估提示词的质量的定量指标和考虑因素：

1) BLEU 分数：BLEU 是一种常用的自然语言处理评估指标，用于度量生成文本与参考文本之间的相似性。可以计算每个生成的响应与参考答案之间的 BLEU 分数，然后取平均值。较高的 BLEU 分数表示更好的匹配。

2) ROUGE 分数：ROUGE 是用于评估文本摘要和机器翻译等任务的指标，也可用于评估生成文本的质量。它可以测量生成文本中的重叠词汇和短语与参考文本之间的相似性。

3) 流畅性(Fluency)：可以使用语言模型来评估生成文本的流畅性。较高质量的生成应该具有更高的概率，而较差的生成可能会被模型视为不够流畅。

4) 相关性(Relevance)：可以使用文本相似性度量来评估生成文本与问题或任务的相关性。例如，可以计算生成文本与问题之间的余弦相似度。

5) 答案完整性(Completeness)：如果问题要求生成文本包括特定信息或答案的各个方面，可以设计指标来检查生成文本是否完整。

6) 多样性(Diversity)：评估生成文本的多样性，确保模型不会反复生成相同或非常相似的答案。可以使用 n-gram 重复度等指标来衡量。

7) 逻辑性(Coherence)：评估生成文本的逻辑性和连贯性，确保答案不包含自相矛盾的信息。

最终的公式可以综合考虑上述指标，权衡它们的重要性，以得出一个总体质量分数。不同任务和应用可能需要不同的公式和权重分配，因此需要根据具体情况进行定制。此外，为了建立可靠的评估体系，需要进行大规模的人工评估和验证，以确定评估指标的有效性和准确性。

以下是一个示例公式{latex 语言}，可以用于评估生成文本的质量：

$$\begin{aligned}\text{Socre} = w_1 \cdot \text{BLEU} + w_2 \cdot \text{ROUGE} + w_3 \cdot \text{Fluency} + w_4 \cdot \text{Relevance} \\ + w_5 \cdot \text{Completeness} + w_6 \cdot \text{Diversity} + w_7 \cdot \text{Coherence}\end{aligned}$$

在这个公式中：1) BLEU、ROUGE、Fluency、Relevance、Completeness、Diversity 和 Coherence 分别代表了各个维度的分数或度量指标。2) (w_1, w_2, w_3, w_4, w_5, w_6, w_7\)是各个维度的权重，用于控制各个维度在总体评分中的重要性。这些权重需要根据任务和应用的需求进行调整和优化，以确保打分公式符合实际情况。

4. 研究结论

4.1. 结论

本研究探讨了如何通过提示词工程的设计和大语言模型(LLM)的能力，实现多种英语口语场景的适配和优化。通过设计不同场景的提示词，我们模拟了雅思口语、K12 应试和学龄前教案设计等多种口语教育和评估场景，以满足不同学习者和教育需求。同时，我们借助百度飞桨开放平台 AISTUDIO、chathub.gg 和提示词 perfect 等平台，进行了实验和对比研究，并提出了提示词评估公式，用于评估设计的提示词在不同场景下的性能和效果。

通过实验和评估，我们得出了如下结论：使用提示词工程的方法可以有效地适配多种英语口语场景，提高口语教育和评估的质量和效率。大语言模型(LLM)在不同口语场景下表现出了良好的潜力，可以根据设计的提示词生成与场景相关的高质量回复。在评估生成文本质量时，需要综合考虑多个指标和因素，以确保生成的回复既准确又流畅。定制化的评分公式可以根据具体任务和应用进行优化，以满足不同需求。

综上所述，本研究为英语口语教育和评估领域的应用提供了一种有效的方法，通过提示词工程的设计和大语言模型的应用，可以实现多种口语场景的适配和优化，为学习者和教育者提供更好的口语学习和评估体验。然而，仍然需要进一步的研究和实验，以进一步提高系统的性能和效果，以满足不断变化的教育需求和技术进展。

4.2. 不足之处

尽管本研究取得了一些积极的成果，但仍存在一些不足之处需要进一步考虑和解决：1) 缺乏进一步量化的指标。在本研究中，我们使用了一系列定量指标来评估生成文本的质量，包括 BLEU 分数、ROUGE 分数等。然而，这些指标仍然无法完全捕捉生成文本的质量，特别是在口语教育和评估领域。未来的研究可以探索更多细化的指标，以更全面地评估生成文本的准确性、流畅性和相关性。2) U 形性能曲线的存在[6]。我们观察到了一个现象，即大语言模型在使用提示词工程时，更擅长使用出现在输入上下文的开头或结尾的相关信息，而过多的提示词工程可能无法完全发挥作用。这导致了一个性能曲线呈 U 形的情况，提示词工程的效果在某个中间点达到最佳。这一现象需要更深入的研究和优化，以提高模型在多种场景下的性能表现。

4.3. 后续研究

在后续研究中，我们可以关注以下方面：1) 提示词工程的扩展应用。提示词工程不仅适用于口语教育和评估领域，还可以扩展到虚拟人、文生图等领域。特别是在英语教育的自动化智能对话阶段(AIGC)，提示词工程为提供高质量、个性化教育体验提供了可行性。未来的研究可以深入探讨提示词工程在不同领域的应用潜力。2) 模型性能的提升。尽管国内的大语言模型在一定程度上能够胜任任务，但与 OPENAI 的 GPT-4 等国际水平的模型相比，仍存在差距。在未来的研究中，我们可以寻求提升国内模型性能的方法，以便更好地支持提示词工程的应用。这可能涉及到模型结构的改进、更多的预训练数据，以及更强大的计算资源的支持。

基金项目

本论文文章由“中国矿业大学(北京)大学生创新训练项目资助(校级项目编号 202308032)”和“中央高校基本科研业务费专项资金资助”资助。

参考文献

[1] 吴砥, 李环, 陈旭. 人工智能通用大模型教育应用影响探析[J]. 开放教育研究, 2023, 29(2): 19-25+45. https://doi.org/10.13966/j.cnki.kfjyyj.2023.02.003

[2] 张辉, 刘鹏, 姜钧译, 曾雄. ChatGPT: 从技术创新到范式革命[J/OL]. 科学学研究, 2023: 1-15. https://doi.org/10.16192/j.cnki.1003-2053.20230626.001, 2023-09-08.

[3] 王婧. 基于联合分析的互联网新产品设计研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2019. https://doi.org/10.27061/d.cnki.ghgdu.2019.006014

[4] 卢宇, 余京蕾, 陈鹏鹤, 李沐云. 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. https://doi.org/10.13541/j.cnki.chinade.20230301.001

[5] 杨锦锋, 梁先桂, 王刘安, 等. 基于 Prompt 策略的医疗对话生成[J]. 中文信息学报, 2023, 37(4): 118-125.

[6] Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F. and Liang, P. (2023) Lost in the Middle: How Language Models Use Long Contexts. https://arxiv.org/abs/2307.03172