

A Lightweight Application of Large Language Model for Vocational Colleges

Yuxiao Wu

Shanghai Technical Institute of Electronics & Information, China

Email: wuyuxiao1991@126.com

Abstract

Generative Artificial Intelligence (Generative AI), represented by ChatGPT, has attracted widespread attention in recent years. This research endeavors to create a lightweight vertical Large Language Model (LLM) application in the field of communication and information engineering education in vocational colleges. By thoroughly analyzing factors such as the convenience, access efficiency, cost, and legality of existing major models, this research has chosen ByteDance's Doubao model and, combined with client and server technologies, built an integrated learning assistant for students in vocational colleges. The client of the application is mainly responsible for friendly interaction with users and realizes user Q&A through the server interface. The server is built based on the classic Springboot framework and calls the general large model API via the HTTP interface. Moreover, optimizations have been made in prompt engineering. Through steps such as design, optimization, and evaluation, it helps the model to accurately respond to students' questions. Ultimately, this application will provide support for students' self-directed learning, information retrieval, homework, and tutoring, and contribute to the development of artificial intelligence applications in vocational education.

Keywords: Large Language Model; Generative Artificial Intelligence; Prompt Engineering; Learning Assistant

一种针对职业院校的大语言模型轻应用的设计

吴欲晓

上海电子信息职业技术学院, 上海 201411

摘要: 在生成式人工智能 (Generative AI) 备受瞩目的背景下, 本研究致力于为职业院校通信与信息工程教育领域打造一款轻量级垂直大语言模型 (LLM) 应用。通过深入剖析现有各大模型的便捷性、接入效率和成本、合法性等方面因素, 选定字节跳动的豆包大模型作为基础模型, 并结合客户端与服务端技术为高职院校学生搭建了一套集成式学习助手-电子小智。应用的客户端主要负责与用户的友好交互, 并通过服务端接口实现用户问答, 服务端将基于经典的 Springboot 搭建框架搭建, 通过 HTTP 接口的方式调用通用大模型 API, 并在提示词工程方面进行了优化, 经设计、优化与评测等环节, 助力模型精准回应学生提问。最终, 此应用将为学生自主学习、信息检索、作业与辅导等提供支持, 并助力于职业教育中人工智能应用发展。

关键词: 大语言模型; 生成式人工智能; 提示词优化; 学习助手

1 引言

职业教育在当今社会发展中扮演着至关重要的角色, 与普通教育共同肩负着培养人才的重任。随着时代的进步, 职业教育正面临着新一轮的改革浪潮, 新的社会需求和技术发展对其提出了更高的要求, 职业教育的改革和发展要求通识教育向职业教育转型, 旨在提升职业教育的现代化水平, 并为促进社会经济发展和提高国家竞争力提供高质量的人才资源^[1]。

AIGC (Artificial Intelligence Generated Content) 是一种基于人工智能技术的内容生成方法, 目前已广泛应用于文本、语音、图像、视频等领域, 也已应用于各类教育的课程教学。现在应用广泛的由 OpenAI 公司

研发的 ChatGPT、国内百度公司研发的文心大模型、字节头条的豆包大模型等都是最具代表性的大语言模型。以 GPT 系列^{[2][3]}为代表的大语言模型（LLMs）为例，极大地改变了自然语言处理（NLP）任务的学术和实践方法。这些任务的例子包括对话查询^[3]、情感分析^[4]和问题回答^[5]。LLMs 还展示出在生成具有语境丰富的对话和展示逻辑推理能力方面的令人印象深刻的能力。事实上，已经证明 LLMs 甚至能够为复杂的算术问题提供准确的推理步骤，在某些情况下超过了人类专家的能力^[6]。

虽然 ChatGPT 等大模型虽发展迅猛、带来新机遇与资源，但高职的学生对其认知和使用存在很大局限，难以发挥优势辅助学习，在调动主动性、获取信息及解决问题上困难重重。教育者因而聚焦于引导学生适应新环境、用好人工智能技术这一关键问题。

本论文将聚焦于这一重要领域，为职业教育领域的人工智能应用提供新的思路和方法，应用 AIGC 的最新技术，设计开发一款针对职业院校的可运行于多客户端的大模型轻应用“电子小智”，该应用结合客户端和服务端技术，为学生提供一个集成式的学习辅助工具,适应职业教育改革的新要求,为他们的学习过程提供有力的帮助，推动职业院校学生学习体验的变革与提升。

2 “电子小智”的设计思路

应用的客户端的核心功能在于呈现直观且美观的页面，并为用户提供友好的交互界面，确保学生能够轻松操作。在客户端的开发设计过程中，UI 的设计以及界面交互行为的设计是关键的开发难点，需要精心考虑如何优化布局、色彩搭配、操作流程等元素，以提升用户体验；而在服务端开发部分，我们将使用经典的 Springboot 框架搭建服务端应用，服务端主要负责为客户端提供接口，实现对大模型 API 的调用，为客户端提供数据支持和服务。对于服务端而言，开发的重点与难点在于大模型 API 的接入，这涉及到与外部服务的通信和数据传输，需要保证接口的稳定性和安全性。

为了使该应用能够为职业院校学生提供最佳的学习辅助效果，我们在大模型的选择上进行了充分调研。对国内外多个经典模型的接入方式进行了细致研究，从多个维度综合考量，包括易用性、费用以及调用速度等。经过全面权衡，我们最终决定选择字节跳动旗下的豆包大模型。该模型在性能、资源可用性和成本效益方面展现出独特优势，更有利于快速便捷实现本应用的开发需求，从而更好地服务于学生的自主学习。

在实施步骤上，我们将按照需求分析、系统设计、开发实现、测试部署等阶段逐步推进。通过合理的系统整体设计、先进的技术路线、科学的研究方法以及严谨的实施步骤，本研究构建了一个功能完善、性能优良的个性化学习轻应用。图 1 展示了具体的研究方法和实施步骤。

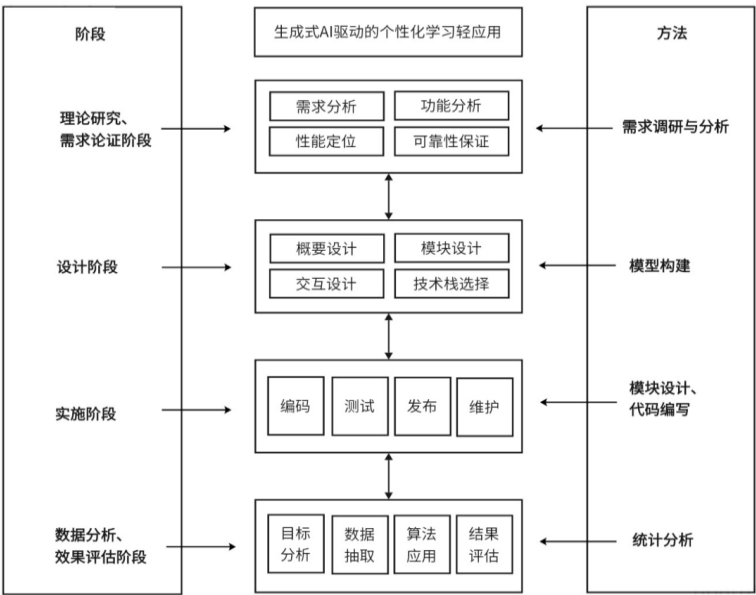


图 1 实施路线图

3 大模型优化的方向与实践

在通用大模型的基础上，研究主要考虑在提示词工程（Prompt Engineering）和有监督精调（Supervised Fine-Tuning，或简称 SFT）两个方面优化大模型的回答。提示词工程无需修改模型参数，而是基于特定的任务指令来增强模型的效果^[7]，而 SFT 是在已有的模型上继续调整参数，来达到和下游任务对齐的目的。通过对这两个方面的实践和效果比较，我们选择了提示词工程作为主要的模型优化方向。

3.1 提示词工程的探究和实践

在自然语言处理（NLP）和对话系统中，提示词（prompt）通常是指用户输入的文本或问题，提示词工程已经成为扩展大语言模型（LLMs）和视觉语言模型（VLMs）能力的不可或缺的技术，它通常可以使人工智能模型能够更好地理解用户的意图和要求，并生成更准确、有用的响应。所以，我们需要探索不同 prompt 对模型输出的影响，创造出能优化模型输出的提示。

一般业务优化的过程，主要包括三个环节，prompt 设计、prompt 优化和 prompt 评测^[8]，如下我们分别分三个部分来进行介绍。

在 Prompt 设计方面，首先需要明确目的，即准备让模型做什么样的事情。在明确目的之后构造 prompt，构造的过程主要包括如下关键点：提供更多提问相关的细节；引导结构和组织要求；限制模型输出格式。

在 Prompt 优化方面，我们在如下几个方向进行探究：让模型扮演一个角色，以使得模型的输出会更符合人类的表达方式，从而更容易被人类理解；提供样例，因为有时指令可能不容易描述清楚，提供示例会更容易；指定完成任务需要的步骤；提供参考内容针对特定领域的问题，帮助模型更好地理解问题的背景和上下文。当然除了以上几点，还有一些其他的优化方向，此处我们暂时只针对相对易于实现的优化方向进行探讨。

在完成上面的 prompt 设计和优化之后，prompt 会发生变更，但是如何判断这个变化是否真的使系统变得更好，通常是先通过看几个例子来确定哪个更好，但是当样本量小时很难区分是真正的改进还是运气导致，所以需要设计一个好的评估集，在经过若干次迭代之后，用评估集来判定系统是否真正得到了改进。这个方面我们暂时不做详细阐述。

3.2 SFT 精调方向的探究与实践

在自然语言处理（NLP）领域，有监督微调（SFT）是一种用来提升大模型在某一特定领域的表现的至关重要的技术手段。SFT 指的是用户提供一份标注好的数据集，即通过给基座模型输入某种特定格式的包含输入的提示词和预期输出的返回的数据集，在已有的模型上继续调整参数，来达到和下游任务对齐的目的。

在进行 SFT 之前，我们需要搞明白什么时候需要 SFT，只有在通过提示词工程无法解决或提示词中描述过于复杂时，才会使用 SFT 精调来调整大模型参数，所以，做 SFT 的前置依赖是把提示词工程做到极致，即通过优化 prompt 已经不能解决剩余的坏用例，特别是当对大模型输出内容有格式要求时，而使用 prompt 优化后，模型仍有部分用例不符合要求，此时才会期望通过 SFT 来减少 prompt 中的内容，加速线上推理的耗时。需要特别注意的是，SFT 并不是对于知识库扩充的一种手段^{[9][10][11]}，只有当在某些特定场景下 SFT 才能最大化地发挥其作用，如以下场景：设置风格、基调、格式或其他需要定性的场景；纠正未能遵循复杂提示的问题；以特定方式处理许多边缘情况；执行难以在提示中阐明的新技能或任务^[12]。

当我们想使用 SFT 进行模型调优时，一开始不需要急着构造大量 SFT 数据集，可以先用少量数据（50 条~100 条）对模型做 SFT 后观察真实评估是否有收益。如果有收益，可以尝试以部分数据为种子数据集继续扩充。如果没有收益，那么再重新检查 SFT 数据集的质量，这部分可以参见 SFT 数据质量判断^[12]。

基于以上的知识，研究进行了小范围的种子数据集的构造和 SFT 精调测试，发现并没有带来明显改善，并且所期望的效果完全可以由提示词优化来实现，所以，SFT 并没有被应用到该模型的优化当中。

4 电子小智的技术实现

4.1 大模型 API 的调用

该应用采用 Springboot 框架搭建服务端，实现对大模型 API 的调用。在开发过程中，大模型 API 的选择是关键环节。在深入调研了包括百度文心大模型、字节头条豆包大模型以及 ChatGPT 等国内外多个知名大模型的 API 接入方式后，从接入便捷性、开发效率、成本和合法性等多维度进行考量，我们选定了字节头条的豆包大模型作为基座模型。

根据 API 接入文档，我们将请求地址设为 `https://ark.cn-beijing.volces.com/api/v3/chat/completions`，并添加一定的`Authorization` 授权头信息，服务端通过创建 HttpClient 实例与大模型 API 通信，执行请求并获取响应后，服务端会检查响应状态码并提取响应体内容，将解析后的模型响应信息整合并返回给客户端，实现与大模型之间的服务调用和信息交互。图 2 展示了本应用的用户-客户端-服务端-大模型 API 的请求时序图。

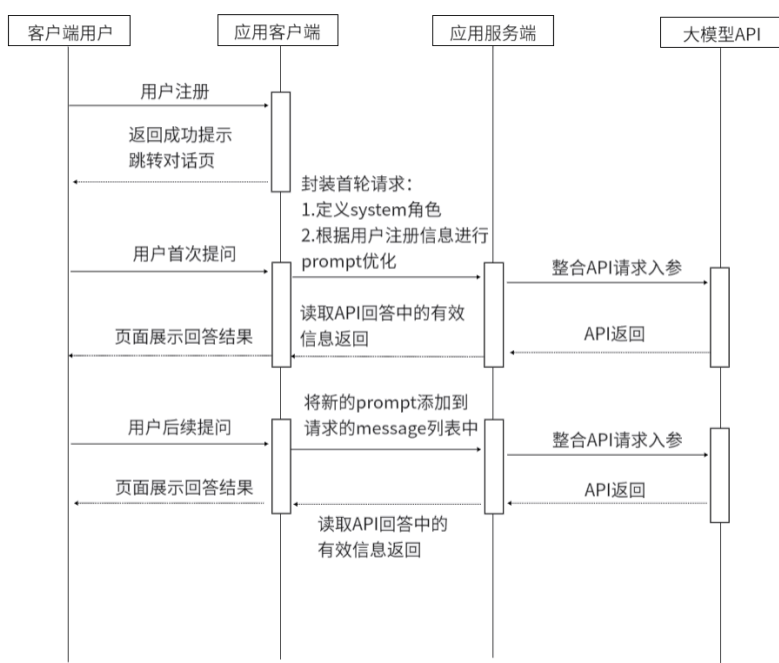


图 2 请求时序图

4.2 客户端开发

在大模型轻应用的开发过程中，客户端软件的构建是至关重要的一环，其核心在于为用户提供便捷且友好的操作界面，并实现与服务端及大模型 API 的有效交互。

该客户端软件主要包含注册页面的开发，要求布局结构清晰，所以需要在 HTML 页面中引入外部的 CSS 文件以美化页面样式，展现良好的视觉效果。注册页面涵盖了多个重要的信息收集区域，包括用户的姓名、专业、年级、学习偏好和学习基础等信息，并通过必填项设置确保录入信息的完整性。页面通过 JavaScript 脚本添加了表单提交的事件监听，当用户点击“注册”按钮提交表单时，若用户已完整填写所有必填信息，这些信息将被持久化存储，用户跳转至应用的主页。

当用户完成注册信息的输入与存储后，这些信息将在后续的提示词工程阶段为大模型提供更丰富的查询细节，辅助大模型更准确地理解用户需求。在用户使用学习助手进行提问时，客户端将以 HTTP 请求的形式将提问信息发送服务端接口，并且针对用户提问进行提示词工程的优化，帮助大模型确定学习助手角色和定位，确保大模型的回答与目标领域的知识和语言习惯相契合。而当大模型 API 接口返回相应的计算结果后，服务端接口会提取返回中的有效信息给客户端，通过这种方式，应用的客户端实现了用户信息的收集、存储、发送请求和展示结果的完整流程。图 3 和图 4 为应用的两张主要的用户交互界面，图 3 为注册页，图 4 为学生与智能助手的对话页，通过简单的交互，学生可以方便地对大模型助手进行提问。



图 3 用户注册页



图 4 用户对话页

4.3 优化点

该轻应用的开发和实现还存在诸多可优化的方面。首先, 信息返回方式的优化至关重要。当前应用采用的是非流式返回, 这在一定程度上会影响用户体验, 因为用户需要等待较长时间才能看到完整的信息。为解决此问题, 考虑将其修改为流式返回, 减少用户的等待时间, 为用户提供更加流畅和高效的使用感受; 其次, 大模型的选择也有待进一步优化, 由于开发时间的限制, 本应用现阶段调用的是通用大模型的 API 接口, 并主要依赖提示词工程来优化返回信息, 该方式存在局限性, 后续计划展开对于包括 Llama 3 在内的流行的开源模型的调研, 并进行本地部署模型, 以此来更方便直观地进行模型参数的调优, 使该轻应用在性能和功能上实现进一步的提升。

5 结论

本文聚焦于职业院校学生的学习需求, 设计并实现了一款基于大模型的轻量级应用程序, 通过对国内外大模型的深入调研, 选定了字节跳动旗下的豆包大模型作为核心, 结合客户端和服务端技术, 构建了一个集

成式的学习辅助工具。在应用开发方面，本文详细介绍了客户端网页、服务端接口和大模型 API 交互的技术细节，展现了完整的开发流程和关键技术点。

本研究不仅为职业院校学生提供了便捷、高效的学习辅助工具，也为职业教育领域的人工智能应用提供了新的思路和方法，有助于推动职业教育迈向新的台阶。

参考文献

- [1] 柯亮亮.基于"1+X"证书制度试点背景下高职院校《路由交换技术》课程课证融合教学改革研究[J].电脑知识与技术: 学术版, 2020, 16(27):4.
- [2] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners Advances in neural information processing systems[J]. 2020, 33:1877-1901.
- [3] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.
- [4] Wankhade M, Rao A C S, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges[J]. Artificial Intelligence Review, 2022, 55(7): 5731-5780.
- [5] Zaib M, Zhang W E, Sheng Q Z, et al. Conversational question answering: A survey[J]. Knowledge and Information Systems, 2022, 64(12): 3151-3195.
- [6] OpenAI: Gpt-4 technical report. arXiv preprint arXiv, 2023:2303.08774.
- [7] Sahoo P, Singh A K, Saha S, et al. A systematic survey of prompt engineering in large language models: Techniques and applications[J]. arXiv preprint arXiv, 2024:2402.07927.
- [8] 火山方舟大模型服务平台 Prompt 最佳实践.<https://www.volcengine.com/docs/82379/1221664>.
- [9] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in neural information processing systems, 2022, 35: 27730-27744.
- [10] Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models[J]. Journal of Machine Learning Research, 2024, 25(70): 1-53.
- [11] Mitra A, Del Corro L, Mahajan S, et al. Orca 2: Teaching small language models how to reason[J]. arXiv preprint arXiv, 2023:2311.11045.
- [12] 火山方舟大模型服务平台 SFT 最佳实践.<https://www.volcengine.com/docs/82379/1221660>.

【作者简介】



吴欲晓（1991-），女，汉族，北京邮电大学硕士，上海电子信息职业技术学院专任教师。

Email: wuyuxiao1991@126.com