● 王翼虎，白海燕，孟旭阳（中国科学技术信息研究所，北京 100038）

大语言模型在图书馆参考咨询服务中的智能化实践探索

摘 要：［目的/意义］大语言模型的爆火为智能问答系统带来了颠覆性变革，给图书馆参考咨询服务的智能化建设提供了创新方向。文章旨在探究大语言模型在图书馆参考咨询服务中的实际应用方案，并评估其生成答案的效果，以期为图书馆的智能化创新发展提供参考。［方法/过程］采用基于p-tuning的大语言模型微调方案提高模型的智能性，根据问答数据构建本地知识库以规范模型内容生成，并利用langchain应用框架构建咨询系统，最后设计评价指标进行主客观综合效果评估。［结果/结论］通过大语言模型微调+langchain本地知识库的联合应用方案，既能发挥模型生成内容的智能性，同时生成内容得到正确规范，生成答案BERT Score的*F*1值达到0.823，验证了其在参考咨询服务中的可行性，为智慧图书馆的AI革新提供创新方向。

关键词：大语言模型；人工智能生成内容；智慧图书馆；参考咨询

DOI：10.16353/j.cnki.1000-7490.2023.08.012

引用格式：王翼虎，白海燕，孟旭阳．大语言模型在图书馆参考咨询服务中的智能化实践探索［J］．情报理论与实践，2023，46（8）：96-103.

An Intelligent Practical Exploration of Large Language Model in Library Reference Consulting Service

Abstract：［Purpose/significance］The explosion of large language models has brought disruptive changes to intelligent question and answer systems and provided innovative directions for the intelligent construction of library reference consulting services. This study aims to investigate the practical application scheme of Big Language Model in library reference consulting service and evaluate its effectiveness in generating answers, in order to provide reference for the development of intelligent innovation in libraries. ［Method/process］A p-tuning-based fine-tuning scheme of the big language model is adopted to improve the intelligence of the model, a local knowledge base is constructed to regulate the model content generation based on the question and answer data, and a consulting system is constructed using the langchain application framework, and finally evaluation indexes are designed to evaluate the comprehensive subjective and objective effects. ［Result/conclusion］The joint application scheme of Large Language Model Fine-tuning + langchain local knowledge base can bring into the intelligence of model-generated content, and at the same time the generated content is correctly standardized. The *F*1 value of generated answer BERT score reaches 0.823, which verifies its feasibility in reference consulting service and provides innovative directions for AI innovation in smart libraries.

Keywords：large language models；artificial intelligence generated content；smart library；reference consulting

0 引言

2022年11月，OpenAI的大语言模型ChatGPT一经上线便火遍全网，成为史上用户数量增长速度最快的消费级应用[1]，而在2023年3月，GPT-4的横空出世更是开启了人工智能的热潮。人工智能生成内容（Artificial Intelligence Generated Content，AIGC）作为大语言模型的应用，为智能问答系统带来了颠覆性变革，使其可以高度智能化地理解人类语言，并生成连贯、自然的对话内容。在此环境下，全球各大企业都在迅速推出自己的大语言模型以及相关应用，如谷歌的BARD[2]、Meta的LLaMA[3]、百度的文心一言[4]和阿里的通义千问[5]等。

智慧图书馆也应在此人工智能的技术机遇下，得到创新式的发展。参考咨询服务作为图书馆最重要的工作之一，其数字化水平是衡量现代图书馆整体服务水平的重要标志[6]，大语言模型优秀的对话能力可以使参考咨询服务的智能性得到快速发展。目前国内图书馆参考咨询系统智能化水平不足，无法对用户的问题做出个性化回答，有学者认为其原因之一是中文自然语言处理技术的应用限制[7]，只能将预设答案进行匹配后原文输出，而借助大语言模型可以解决行业智能咨询需求，为图书馆用户提供更具真实感、沉浸感的体验[8]，从而实现图书馆咨询服务在智能技术发展上的“大跃进”[9]。因此在图书馆参考咨询服务中应用大语言模型具有重要的实践意义，本文首先采

用一套基于 p-tuning 模型微调 + langchain 本地知识库的结合方案，实现大语言模型在参考咨询服务中的应用实践，并且针对大语言模型生成内容的特点设计评价指标，进行主客观综合效果评估，验证可行性及应用意义；同时设想未来改进方向，为图书馆参考咨询服务提供了新的思路和方法，帮助图书馆更好地适应信息时代的变化，提供更加创新和有价值的服务。

1 相关研究

1.1 参考咨询服务发展概况

参考咨询工作是图书馆中最为重要的工作之一，是图书馆员对读者在利用文献和寻求信息方面提供帮助的工作。关于国内外将人工智能应用于图书馆的研究由来已久，在 2004 年汉堡大学图书馆系统就已设计并开发了全天候开放的 Stella[10]，之后美国内布拉斯加大学林肯分校图书馆也推出了 Pixel[11] 等。国家科技图书文献中心（NSTL）是我国最早推出准实时参考咨询服务的图书馆之一[12]，之后还有清华大学图书馆的“小图”[13] 以及上海交通大学图书馆 IM 咨询机器人[14] 等。

随着自然语言处理技术的发展，图书馆参考咨询服务的构建技术也得到了不断改进。李记旭[15] 于 2009 年提出基于范例推理的数字参考咨询系统，其主要思想是把待求解的问题与系统中已存的范例进行类比分析，找出与待求解问题最相似的范例。2012 年，李文江等[16] 提出了基于 AIMLBot 的实时虚拟参考咨询服务，人工智能标记语言（Artificial Intelligence Markup Language，AIML）是一种基于 XML 标准的丰富标签库，AIML 具备一定的学习能力但仍存在智能性不足的问题。2019 年，陆伟等[17] 在图数据库中进行子图匹配计算相似度实现答案的生成，知识图谱开始应用于参考咨询服务，施国良等[18] 认为还可以对其他图结构向量化，比如对节点和路径进行向量化，通过向量间相似度实现知识问答。同年，朱娜娜等[19] 提出一种融合人物画像的对话生成模型，使用基于模板方式和基于 Bootstrapping 的机器学习方式对人物画像进行建模，提升了数字参考咨询问答的智能化程度。2022 年，刘泽等[20] 基于深度学习和文本匹配等技术构建多策略混合问答系统模型，采用了 Bi-LSTM 和 CNN 结合的深度学习模型提升问题匹配精确性。

图书馆参考咨询系统的智能化改进是发展目标和未来趋势，由于实际场景中用户遇到的问题多样化，而向咨询系统提问的方式也呈现多样化，如果系统智能化不足则会难以理解用户的想法，从而只能转向人工服务，加大图书馆咨询员的工作量。

1.2 大语言模型在图书馆的应用概况

2018 年，谷歌发布了一款基于 Transformer 的预训练模型 BERT[21]，标志着人工智能进入了预训练模型的时代，并引发了 AIGC 技术能力的质变。通过微调（Fine-tune）预训练模型，可以解决前期基础模型使用门槛高、训练成本高、内容生成简单等问题。GPT 全称为“Generative Pre-Trained Transformer”，即生成预训练语言模型，大型语言模型（Large Language Model，LLM）是 GPT 类技术所依赖的核心模型，它采用深度学习技术，通过预训练学习大量自然语言文本，然后通过自注意力机制和 Transformer 结构实现生成文章、回答问题、翻译等自然语言处理任务。

大语言模型的热潮，也引起了包括图书馆在内各行各业的密切关注。王启云[22] 认为信息技术的发展使图书馆产生了许多颠覆性变化，而 GPT 类技术可以看作是功能强大的数字参考咨询服务，可以有效解答许多问题。也有学者认为 GPT 类技术可以促进图书馆聊天机器人系统地发展，加快图书馆整体服务流程，使用户咨询的问题得到快速响应，提升服务效率[23]，并且其情景理解能力和启发性内容生成能力，也可以使咨询服务更具智能化[24]。除此之外还有学者在交互类层面上提出应用分析，赵瑞雪等[25] 认为图书馆应该深化大语言模型应用，如文献数据加工以及知识服务等，以此提升图书馆智能化加工和服务程度。张慧[26] 认为由 GPT 类技术可以打造无人值守的图书馆系统，实现全天候高效运转，驱动智慧图书馆实现更加智能化、多元化、交互式的图书馆体验，给用户带来更加高效、个性化的服务。

综上所述，在如今以 GPT 类技术和大语言模型为主的技术革新时刻，图书馆参考咨询服务也将获得创新发展的机会。如何充分发挥大语言模型的优势，制定一套切实可行的方案提升现有参考咨询服务的智能化水平成为研究的难点，本研究以此作为目标进行实验并制定方案。

2 大语言模型应用方案设计

基于前述分析，在充分发挥大语言模型优势的同时，为规范其答案生成保证足够完整和准确，本文采用一种大语言模型微调 + 本地知识库联合使用方案，详细方案包括：以清华大学开源模型 ChatGLM-6B 为实验模型，首先进行基于 P-Tuning v2 的大语言模型微调，再将微调后的模型接入基于 Langchain 的框架，使用现有参考咨询数据集生成本地知识库，通过 Sentence-transformer 向量检索的方法筛选出 Top k 条与用户提问最接近的答案作为参考依据，整体生成 Prompt 输入模型，通过大语言模型的理解分析，依照参考生成最终答案，并且如果提问超出范围的

问题，禁止模型自动生成答案从而误导用户，具体方案设计如图1所示。

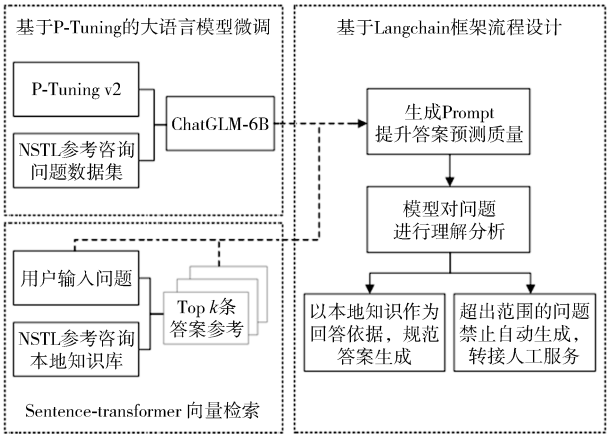

图1 系统架构图

Fig. 1 System Architecture Diagram

2.1 大语言模型选型

大语言模型爆火以来，国内外各大研究机构推出了各种开源大语言模型，部分较为常用的开源模型如表1所示。

表1 国内外主要开源大语言模型

Tab. 1 Open source large language models

模型名称	研究单位	参数量	训练数据量
LLaMA[3]	Meta	70亿、130亿、330亿、650亿4种参数规模	1.4万亿token
Alpaca[27]	斯坦福大学	70亿	5.2万条问答指令数据
Koala[28]	加利福尼亚大学伯克利分校	130亿	50万条高质量数据集合
鹏程·盘古α[29]	鹏程实验室、华为	26亿、130亿和2000亿共3种参数规模	1.1TB高质量中文语料数据集
ChatGLM[30-31]	清华大学KEG实验室和智谱AI	60亿和1300亿共2种参数规模	1TB中英比例为1∶1的token

表1中的Alpaca和Koala都是在LLaMA的基础上使用英文语料进行微调构建的，目前国内的开源大语言模型主要是华为和鹏程实验室主导的鹏程·盘古α，以及清华大学主导的ChatGLM。

鹏程·盘古α是由以鹏城实验室为首的技术团队联合攻关，训练出业界首个2000亿参数以中文为核心的预训练生成语言模型，在知识问答、知识检索、知识推理、阅读理解等文本生成领域表现突出，具备很强的小样本学习能力。

ChatGLM是由清华大学推出的一个支持中英双语的开源大语言模型，其中ChatGLM-6B具有62亿的参数，通过模型量化技术，可以让用户在消费级的显卡上进行部署。ChatGLM-6B使用了和ChatGPT相似的技术，并且针对中文问答和对话进行了优化。除此之外，通过监督微调（Supervised Fine-Tuning）、反馈自助（Feedback Bootstrap）、人类反馈强化学习（Reinforcement Learning from Human Feedback）等方式，使模型初具理解人类指令意图的能力。

同时ChatGLM-6B推出了基于P-Tuning-v2[32]的高效参数微调，可以将需要微调的参数量减少到原来的0.1%，再通过模型量化、Gradient Checkpoint等方法，可以大幅度降低模型在使用中的内存和计算资源消耗，在提升训练效率的同时保持模型的性能不受影响。

由于本文主要使用中文参考咨询数据进行实验，综合考虑中文语料模型效果以及硬件配置要求，故采用ChatGLM-6B作为实验模型。

2.2 基于langchain框架的本地知识库构建

langchain是一个用于构建基于大型语言模型应用程序的框架，大语言模型生成答案往往由其自由发挥，具有不确定性和不准确性等特点，而langchain开发框架针对此问题融合了各种功能，旨在让开发人员构建定制版大型语言模型对话产品，构建本地知识库的具体流程如图2所示。

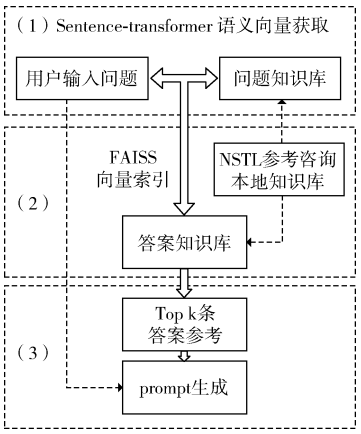

图2 本地知识库构建流程图

Fig. 2 Schematic diagram of the knowledge base construction process

1）首先使用Python语言提取NSTL参考咨询问答数据集中的官方回答，以及用户手册中常见问题构建本地知识库，并将其分为问题知识库和答案知识库。分别获取用户输入问题和问题知识库的句向量，便于后续进行语义匹配，本文使用Sentence Transformers库中的text2vec-large-chinese模型进行句向量的获取，相较于Word2Vec、glove等传统静态词向量模型，Transformers[33]可以获得不同情

景下文本的语义向量。

2）知识库的语义向量获取后，如果对问题逐个相似度计算会造成计算资源浪费、降低回答效率，所以采用FAISS库进行相似度搜索。FAISS[34]是Facebook AI Research团队开发的一款高效的相似度搜索库，提供了多种向量索引算法，可以用于大规模向量数据的处理。向量索引是FAISS的核心概念之一，其中使用倒排索引的方法可以大幅度加快在大规模相似度搜索的速度，其基本原理为：通过将大量向量数据聚类分为多个小组，每个小组分配多个相似向量，正式搜索时先进行向量小组之间的运算，以此避免大多数错误向量的计算，之后再对小组内的向量逐一计算，从而提升整体运算效率，原理如图3所示。

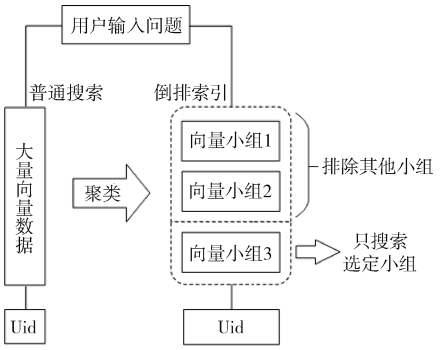

图3　倒排索引原理图

Fig. 3　Inverted File System schematic

3）将相似度搜索后得到的Top k 条答案参考和用户输入问题利用langchain的prompt_template功能生成大语言模型输入需要的prompt（提示），引导大语言模型从答案参考中进行理解分析并生成准确完整的答案，同时强调无法得出答案时禁止自行生成答案，应回答道歉语句并转接人工服务。

综上所述，大语言模型具有高度智能性的特点，模型微调的过程相当于让其明确自身职业和定位，以及对相关专业问题如何引导解答；构建本地知识库的过程相当于规范答案的正确性，约束其联想和思维发散能力，让其在拥有自己理解的基础上参考合乎规则的答案，保证最后结果的有效性。

3　实验与效果分析

3.1　实验环境与参数设置

本实验环境配置为：CPU，12 vCPU Intel（R）Xeon(R) Platinum 8255C CPU @ 2.50GHz；GPU，RTX3090；显存：24GB；Python版本，3.8.16；Cuda版本，11.2。实验超参数设置如表2所示。

表2　实验主要超参数设置

Tab. 2　Experiment main hyperparameter settings

超参数	中文解释	数值
max_source_length	最大输入序列长度	256
max_target_length	最大输出序列长度	256
train_batch_size	每批次训练数据量大小	1
learning_rate	学习率	2e－2
max_steps	最大训练步数	1000

3.2　数据来源与模型训练

使用Python语言构建NSTL参考咨询数据集，内容包括参考咨询常见问题以及用户手册的内容，共得到数据968条，对问题数据进一步分析，得出具体问题类型分布，如表3所示。

表3　数据集问题类型示例

Tab. 3　Example of dataset problem type

问题类型	类型介绍	问题举例	问题数量
介绍问题	此类问题包括咨询服务的身份职责，以及各领域门户和成员单位的介绍	1. 请问你是谁？ 2. NSTL参考咨询方式？ 3. 领域门户是什么？	134
账号问题	此类问题包括登录异常、密码找回、个人与集团用户注册等账号相关咨询	1. 为什么登录不了？ 2. 密码忘了怎么办？ 3. 集团用户如何注册？	77
费用问题	此类问题包括收费标准、支付方式、优惠政策、退款申请等费用相关咨询	1. 代查代接如何收费？ 2. 支持什么方式支付？ 3. 申请单费用有优惠吗？	150
使用问题	此类问题包括NSTL各种功能的使用方式和引导	1. 如何进行快捷检索？ 2. 检索不到如何索取原文？ 3. 文献号如何解读？	173
服务问题	此类问题包括向NSTL申请各类服务时产生的相关问题	1. 图片能否给我彩扫？ 2. 邮件附件损坏请重发 3. 申请变动机构IP地址	194
资源问题	此类问题包括对各类资源以及数据库等情况的咨询	1. 全文开通都有哪些资源？ 2. 如何查询EI索引？ 3. SCI数据库你们有吗？	44
联系方式问题	此类问题包括各领域门户及成员单位的具体负责人以及网址等联系方式	1. 领域信息门户联系方式？ 2. 南京服务站网址是什么？ 3. 地学服务站有问题该找谁？	196

根据如上数据集与环境配置进行模型训练，并记录每轮（Epoch）的训练损失值（Training Loss）和学习率（Learning Rate）进行模型训练效果的评价，最终结果如图4所示。

由图4可知，随着训练轮数的增加，训练损失值逐渐减小，代表模型训练效果逐渐提高；同时学习率不断降

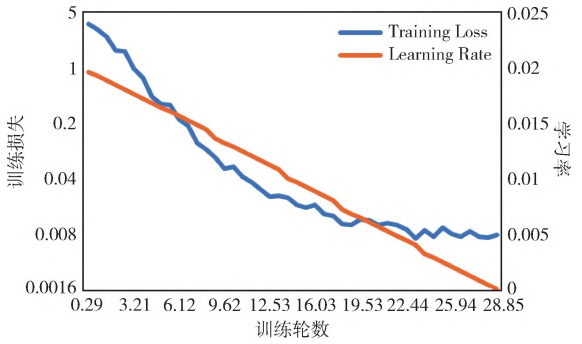

图4 模型训练损失与学习率

Fig. 4 Model training loss and learning rate

低，说明模型参数逐渐接近最优解，模型效果趋于稳定。使用训练好的参数进行问答实验，通过 gradio 库构建网络可视化问答聊天框，问答测试示意图如图5所示。

图5 可视化参考咨询问答测试示意图

Fig. 5 Visualization reference consulting Q&A test schematic

3.3 实验结果与分析

目前国内外大语言模型应用的主流方案是利用大规模现实文本进行模型微调，如法律领域的 LawGPT[35]，以及医学领域的 HuatuoGPT[36] 和 DoctorGLM[37]。但参考咨询服务重点在于帮助用户完成图书馆的各项服务，而每种服务均有明确、固定的流程，因此对答案的严谨性具有更高要求。为测试本文方案的实际效果，本文将分别测试使用主流方案“仅微调”策略以及“微调+知识库”结合策略，依据表3中的分类结果，对每个类别进行测试，对比两种方案的性能差异并进行分析。

由于大语言模型生成内容具有智能性、随机性等特点，所以传统基于词重叠的评价指标并不能很好地描述模型效果，本文采用基于语言模型的评价指标 BERT Score[38] 进行评估，该指标使用了上下文嵌入计算标识的相似性，通过 BERT 模型分别对生成文本和参考文本转化为 token 提取特征，再计算两段文本中每一个词对应的内积，以此构造一个相似性矩阵，通过计算候选句子与参考句子中每个标记的相似性得出最终分数，计算示意图如图6所示。

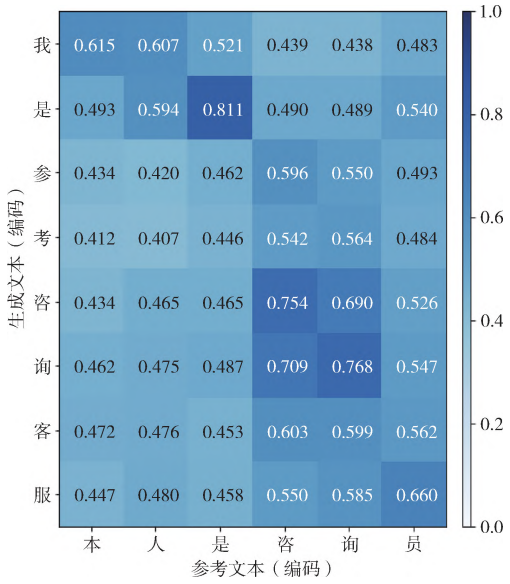

图6 相似度矩阵示意图

Fig. 6 Similarity Matrix schematic

基于此矩阵计算两段文本的最大相似性得分并归一化，最终得到准确率（Precision）、召回率（Recall）和 $F1$ 值如表4所示。

表4 模型效果评价结果

Tab. 4 Model effect evaluation results

问题类型	仅模型微调			模型微调+本地知识库		
	Precision	Recall	$F1$	Precision	Recall	$F1$
介绍问题	0.831	0.839	0.835	0.797	0.882	0.837
账号问题	0.637	0.583	0.608	0.888	0.901	0.894
费用问题	0.702	0.664	0.682	0.846	0.843	0.844
使用问题	0.570	0.542	0.555	0.712	0.799	0.753
服务问题	0.643	0.617	0.629	0.825	0.831	0.817
资源问题	0.619	0.527	0.570	0.935	0.928	0.932
联系方式问题	0.800	0.659	0.723	0.840	0.879	0.859
总计	0.686	0.633	0.657	0.835	0.866	0.848

由表4可知，仅进行模型微调时只在介绍问题上有较明显效果，整体 $F1$ 值达到0.835，但在图书馆具体服务类问题上表现较差，说明进行模型微调可以回答一些较为笼统的介绍型问题，但难以对具体任务的流程细节进行准确生成，总体 $F1$ 值仅有0.657。

本地知识库的加入让系统有了更明确的信息来源，使整体效果有明确提升，尤其在资源、联系方式等有固定答案的问题上表现较好。但在具体使用问题上 $F1$ 值仅得到0.753，可能由于实际问题的多样化会引起系统理解偏差

等情况，但总体 *F*1 值达到 0.848，已取得不错的效果。

通过评价情况可以得出，使用大语言模型微调 + 本地知识库结合的方案具有在小规模数据集上生成更严谨答案的能力。采用 BERT Score 的客观指标可以对生成内容进行整体评估，但与以往智能问答系统不同，大语言模型的生成内容具有更高的智能性和联想性，本文将在实例分析中针对大语言模型生成内容的特点进行深入分析。

3.4 实例分析

为进一步分析大语言模型生成内容效果，本文从各种类型中挑选最具代表性的问题进行实例分析，具体测试结果如表 5 所示，为便于后续深入研究，将测试结果中与答案完全相符部分标记为蓝色（计算机显示颜色，下同）、错误部分标记为红色、模型联想生成部分标为下划线显示。

表 5 系统结果测试表

Tab. 5 System results test table

问题类型	问题	仅模型微调	模型微调 + 知识库	参考答案
介绍问题	请问你是谁？	您好！我是参考咨询客服，很高兴为您服务！有什么能帮到您吗？	您好！我是参考咨询客服，很高兴为您服务！如果您有关于图书馆参考咨询的问题，欢迎您前来咨询，我会尽力为您解答	您好！我是参考咨询客服，很高兴为您服务！有什么能帮到您吗？
账号问题	密码错误了怎么办	密码错误，请拨打参考咨询的电话，请求协助解决	密码错误了，可以在主页右上角—登录—（登录框下方）忘记密码；您可凭“手机号”或者“用户名 + 邮箱”获取新密码。（答案不全）	如果绑定了手机，可使用“随机密码登录”，用手机号和验证码即可登录； 主页右上角—登录—（登录框下方）忘记密码；您可凭“手机号”或者“用户名 + 邮箱”寻回密码（也需要绑定手机或邮箱）
费用问题	支付方式都有哪些	支付方式包括网上付费、电话付费、电子邮件支付、刷卡支付等。请见主页右上角“帮助”，“注册与付费”，进入收费标准，有详细的支付方式说明	支付方式包括网上支付、邮局汇款、银行转账、直接交费四种方式。网上支付包含支付宝和各类银行卡，邮局汇款、银行转帐、直接交费三种方式。在导航栏“帮助”栏目下有详细的说明	北京中心站包含网上支付（包含支付宝和各类银行卡），邮局汇款，银行转账，直接交费四种方式。 服务站用户请直接联系您所在的各服务站点
使用问题	怎么进行快捷检索？	快捷检索需要注册账号，请见主页右上角“帮助”，“注册与付费”，进入收费标准，有详细的注册说明。 然后，在快捷检索页面，勾选要检索的文献类型，点击检索，即可快速检索到需要的文献。 如果需要下载文献，点击文献页面右上角“下载”可下载全文。 快捷检索可以帮助用户更快速地找到需要的内容，方便实用	通过导航条第二条“文献检索”——进入，键入关键词或主题词，然后选择相应的中文或西文数据库，点击检索即可。 快捷检索位置明显，可实现快速检索，默认的文献类型为中外文期刊、会议、学位论文，其他文献类型可选。高级检索：提供专业检索。撤稿申请需向服务中心提供 1 作者的撤稿声明和 2 邮箱地址，发至 service@ nstl. gov. cn。（识别其他答案）	主页快捷检索数据库，是以文献类型来分的，包括中英文：期刊、会议、学位论文、科技报告、专利、文集、图书、标准、计量规程等。默认检索为：中外文期刊、会议、学位论文
服务问题	机构 ip 地址变更流程是什么	机构 IP 地址变更申请，需要先注册成为文献传递用户，然后提交申请，等待服务老师处理。 具体流程： 1. 注册成为文献传递用户； 2. 填写机构 IP 地址变更申请（上传变更前后的网络拓扑图）； 3. 付费； 4. 提交申请； 5. 服务老师处理申请； 6. 成功接收文献或完成付款	机构 IP 地址变更流程： 1. 下载《机构信息变更表》； 2. 签盖《机构信息变更表》； 3. 将 PDF 扫描件和 word 版电子表一并发送至：service@ nstl. gov. cn； 4. 原件寄至：北京市海淀区复兴路 15 号 NSTL 用户服务中心曾老师（收）电话：010-58882323	对 IP 地址发生变化的单位，NSTL 将通知各出版商、代理服务商和 NSTL 网络管理中心，由上述各有关方面在每个双月的月底统一修改有关网络数据库为相关单位开通的 IP 地址配置。 用户需要做的是：下载《机构信息变更表》，签盖完成以后，请完成以下两个步骤： 将 PDF 扫描件和 word 版电子表一并发送至：service@ nstl. gov. cn。 原件寄至：北京市海淀区复兴路 15 号 NSTL 用户服务中心曾老师（收）电话：010-58882323
资源问题	可以提供 sci 的检索服务吗	SCI 检索需要预付费，所以请注册成为文献传递用户后再提交申请	可以的，NSTL 不提供 SCI、EI 等数据库以及出具证明的服务，但个别成员单位在自身服务中可以提供该服务。拥有 SCI、EI 等数据库以及出具证明的单位有：中国科学院文献情报中心、中国科学技术信息研究所、中国医学科学院医学信息研究所。用户可以到上述单位亲自查询，或委托该单位的工作人员检索	NSTL 不提供该项服务，但 NSTL 个别成员单位在自身服务中可以提供该服务。拥有 SCI、EI 等数据库以及出具证明的单位有：中国科学院文献情报中心、中国科学技术信息研究所电子阅览室中国医学科学院医学信息研究所查新咨询室。用户可以到上述单位亲自查询，或委托该单位的工作人员检索
联系方式问题	农业立体污染防治领域信息门户的联系人是谁	领域门户联系人：王俊林电话：010-82620575	农业立体污染防治领域信息门户的联系人是季雪婧，电话是 010-82109652-803，E-mail 是 jixuejing@ caas. cn	领域门户联系人电话（010）E-mail 农业立体污染防治季雪婧 010-82109652-803 jixuejing@ caas. cn
错误问题	可以给我一本高等数学吗	书籍传送需要预付费，不能直接提交申请（代查代借需要预付费再提交，原文传递可以在订购书籍最后一步交费）。 服务站用户请联系各服务站点服务老师拿取	不好意思，作为一个 AI 咨询员我暂时无法回答您的问题，我可以代查代借，如果您需要提高分辨率或彩色扫描，请通过代查代借申请，并备注您的需求。如果文献不清楚请告知您的文献号和联系方式，我们核查后再与您联系。谢谢！ 代查代借查询：010-58882057，service @ nstl. gov. cn	错误问题无答案

由表5中测试内容可以得出以下结论：

本方案回答更具人性化、内容更严谨。在服务问题中，模型直接针对问题回答用户所需要做的内容，同时将所有步骤整理成顺序编号的流程，将下载和签盖步骤也进行了总结，修改总结内容合理且具有条理；在账号、费用和资源等问题中，模型也会先针对性回答用户的问题，再将相关内容详细解释，更具有内容交互性；不仅如此，测试中特别设置了在知识库中不包含的错误问题，模型十分正确地表示了无法回答，并且给出了自己的理解建议，整体答案属于合理范围，展现了大语言模型的智能性。

模型生成答案存在少量理解错误的情况。在费用问题中，模型将网上支付括号内的内容进行了自行概况和解释，但是误将其他几种支付方式同时概括，出现了内容重复、混乱的情况；在使用问题中，模型擅自将其他文献类型也并入默认检索类型，并且将撤稿申请相关内容也作为答案输出，造成理解错误的情况。在账号问题中，模型认为密码错误应该找回密码，但是忽略了可以用手机号登陆的信息，笔者认为原因在于模型认为问题的重点在于“密码错误”而不在“无法登陆”，属于理解错误，因此造成了答案不全的情况。

为进一步评价大语言模型生成答案的效果，针对大语言模型生成内容常常会根据其理解自行增添或总结内容的特点，本文设计出“内容修改率”（Content Modification Percentage，CMP）和“内容修改合理度”（Content Modification Reliability，CMR）两个主观指标对模型自行理解部分进行测试，其计算公式见公式（1）和公式（2）：

$$CMP = \begin{cases} \dfrac{L(A - A^*)}{|A|} & L(A) > L(A^*)\text{（生成内容）} \\ \dfrac{L(A^* - A)}{|A^*|} & L(A) < L(A^*)\text{（总结内容）} \end{cases} \quad (1)$$

$$CMR = \frac{L(\text{合理内容长度})}{|L(A - A^*)|} \quad (2)$$

式中，生成答案文本为A，参考答案文本为A^*。

由公式（1）可知，当模型生成内容为增添内容时，$L(A - A^*)$代表生成答案中比参考答案多出的内容；当模型生成内容为总结内容时，$L(A - A^*)$代表参考答案比生成答案多出的内容。通过CMP可以评价模型对内容进行自行修改的数量程度。

由公式（2）可知，$|L(A - A^*)|$代表模型总共修改的内容，CMR表示模型生成内容中合理内容的占比，可以评价模型模型对内容进行自行修改的质量程度，对模型生成答案的具体分析情况如表6所示。

由表6中数据可知，仅进行模型微调时，模型自行生成内容达到0.786，但其中只有0.391的内容较为合理，大部分内容并不符合要求。在引入知识库后，模型的CMP降至0.439，说明模型自行生成内容以及对参考答案修改的内容都大幅降低，并且CMR提升至0.871，代表模型生成内容中约有87%的内容是符合正确规则的、合理的修改。

表6 主观指标评价分析表

Tab.6 Subjective index evaluation analysis table

模型策略	仅模型微调		模型微调+知识库	
主观指标	CMP	CMR	CMP	CMR
总计	0.786	0.391	0.439	0.871

综上所述，目前主流的微调方案难以在实际参考咨询服务中进行应用，虽然微调后的模型具备准确理解用户提问的能力，但在答案生成上缺乏咨询服务需求的严谨性。通过本方案可以限制模型自由生成能力，同时保证答案的严谨性和准确性，并且模型能够根据用户的提问对回答内容酌情修改，与传统的将设定答案直接返回相比，有效提升了咨询服务的智能性、交互性以及人性化。

4 总结与展望

本文通过构建一套基于大语言模型微调+本地知识库的结合方案实现了在参考咨询服务的应用实践，并针对性设计评价指标验证其可行性及应用意义，通过与传统方案对比得出本研究优点在于：①内容更具参考咨询服务所需的严谨性。充分利用大语言模型智能性的同时规范其自由生成能力，回答内容不仅更具人性化而且符合具体规定。②具有实际应用的可行性。本文采用支持量化技术的ChatGLM-6B模型，在消费级显卡上即可部署，并采用倒排索引提升系统召回效率，具备实际应用的可行性。

本文存在一些不足，尽管已经通过prompt避免过度理解的情况，但实际应用时仍存在少量错误内容，在今后的研究中可以采用数据集扩充技术增大规模，并根据实际用户的咨询内容增添问题多样性，加强模型的泛化能力，同时加强模型对图书馆咨询业务的具体理解，以此研究如何更好规避大语言模型的过度理解。如今大语言模型因其强大的性能已成为全世界的研究热点，本文期望为图书馆参考咨询服务的创新建设提供参考，通过实践认识大语言模型为图书馆带来的机遇和挑战，积极利用人工智能技术发展智慧图书馆。□

参考文献

[1] DENNEAN K，GANTORI S，LIMAS D，et al. Let's chat about ChatGPT [EB/OL]. [2023-05-05]. https://www.ubs.com/global/en/wealth-management/our-approach/market-news/article.1585717.html.

[2] Google. Google responds to OpenAI with its own chatbot-'bard' [EB/OL]. [2023-05-05]. https://www.washing-

tonpost. com/technology/2023/02/06/google-bard-chatbot.
[3] TOUVRON H，LAVRIL T，IZACARD G，et al. Llama：open and efficient foundation language models［J］. arXiv preprint arXiv：2302. 13971，2023.
[4] 百度. 文心大模型［EB/OL］.［2023-05-06］. https://wenxin. baidu. com/.
[5] 阿里. 通义千问［EB/OL］.［2023-05-06］. https://tongyi. aliyun. com/.
[6] 周泰冰，刘文云. 我国公共图书馆数字参考咨询服务比较分析［J］. 情报理论与实践，2010（12）：84-87.
[7] 李书宁，刘一鸣. ChatGPT 类智能对话工具兴起对图书馆行业的机遇与挑战［J］. 图书馆论坛，2023，43（5）：104-110.
[8] 储节旺，杜秀秀，李佳轩. 人工智能生成内容对智慧图书馆服务的冲击及应用展望［J］. 情报理论与实践，2023，46（5）：6-13.
[9] 郭亚军，郭一若，李帅，冯思倩. ChatGPT 赋能图书馆智慧服务：特征、场景与路径［J/OL］. 图书馆建设：1-16［2023-05-09］. http://kns. cnki. net/kcms/detail/23. 1331. G2. 20230406. 1553. 004. html.
[10] MCNEAL M，NEWYEAR D. Chatbots：automating reference in public libraries［M］. Pennsylvania：IGI Global，2013：101-114.
[11] ALLISON D A. Chatbots in the Library：is it time?［J］. Library Hi Tech，2012，30（1）：95-107.
[12] 顾德南. NSTL 数字化参考咨询服务初探［J］. 图书情报工作，2004，48（1）：19.
[13] 姚飞，张成昱，陈武. 清华智能聊天机器人“小图”的移动应用［J］. 数据分析与知识发现，2014（7）：120-126.
[14] 孙翌，李鲍，曲建峰. 图书馆智能化 IM 咨询机器人的设计与实现［J］. 数据分析与知识发现，2011（5）：88-92.
[15] 李记旭. 基于范例推理的数字参考咨询系统实现初探［J］. 情报理论与实践，2009（6）：78-80.
[16] 李文江，陈诗琴. AIMLBot 智能机器人在实时虚拟参考咨询中的应用［J］. 现代图书情报技术，2012，28（7）：127-132.
[17] 陆伟，戚越，胡潇戈，等. 图书馆自动问答系统的设计与实现［J］. 情报工程，2019，5（2）：5-16.
[18] 施国良，谢泽宇，杨小莉. 高校图书馆复杂网络构建与智慧化应用探索［J］. 图书情报工作，2019，63（23）：106-112.
[19] 朱娜娜，景东，张智钧. 面向图书馆数字参考咨询的人机对话模型［J］. 图书情报工作，2019，63（6）：5-11.
[20] 刘泽，徐潇洁，邵波. 基于多策略混合问答系统模型的图书馆咨询机器人的设计与应用［J］. 新世纪图书馆，2022，309（5）：43-49.
[21] DEVLIN J，CHANG Mingwei，LEE K，et al. Bert：pre-training of deep bidirectional transformers for language understanding［J］. arXiv preprint arXiv：1810. 04805，2018.
[22] 王启云. ChatGPT 对图书馆工作的影响——圃人堂专题讨论综述［J］. 大学图书情报学刊，2023，41（2）：3-9.
[23] PANDA S，KAUR N. Exploring the viability of ChatGPT as an alternative to traditional chatbot systems in library and information centers［J］. Library Hi Tech News，2023，40（3）：22-25.
[24] CHEN Xiaotian. ChatGPT and its possible impact on library reference services［J］. Internet Reference Services Quarterly，2023，27（2）：121-129.
[25] 赵瑞雪，黄永文，马玮璐，等. ChatGPT 对图书馆智能知识服务的启示与思考［J］. 农业图书情报学报，2023，35（1）：29-38.
[26] 张慧，佟彤，叶鹰. AI 2.0 时代智慧图书馆的 GPT 技术驱动创新［J/OL］. 图书馆杂志：1-7［2023-05-10］. http://kns. cnki. net/kcms/detail/31. 1108. G2. 20230411. 1939. 002. html.
[27] TAORI R，GULRAJANI I，ZHANG Tianyi，et al. Stanford alpaca：An instruction-following llama model［J］. GitHub Repository，2023.
[28] GENG Xinyang，GUDIBANDE A，LIU Hao，et al. Koala：a dialogue model for academic research［EB/OL］.［2023-05-05］. https://bair. berkeley. edu/blog/2023/04/03/koala/.
[29] 2000 亿开源中文预训练语言模型「鹏程·盘古 α」［EB/OL］.［2023-05-12］. https://openi. pcl. ac. cn/PCL-Platform. Intelligence/PanGu-Alpha.
[30] ZENG Aohan，LIU Xiao，DU Zhengxiao，et al. Glm-130b：an open bilingual pre-trained model［J］. arXiv preprint arXiv：2210. 02414，2022.
[31] DU Zhengxiao，QIAN Yujie，LIU Xiao，et al. GLM：General language model pretraining with autoregressive blank infilling［C］//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics（Volume 1：Long Papers），2022：320-335.
[32] LIU Xiao，JI Kaixuan，FU Yicheng，et al. P-tuning v2：Prompt tuning can be comparable to fine-tuning universally across scales and tasks［J］. arXiv preprint arXiv：2110. 07602，2021.
[33] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need［J］. arXiv e-prints，2017：arXiv：1706. 03762.
[34] JOHNSON J，DOUZE M，JÉGOU H. Billion-scale similarity search with gpus［J］. IEEE Transactions on Big Data，2019，7（3）：535-547.
[35] LIU Hongcheng，LIAO Yusheng，MENG Yuhao，et al. LawGPT：中文法律对话语言模型［EB/OL］.［2023-06-06］. https://github. com/LiuHC0428/LAW-GPT.
[36] ZHANG Hongbo，CHEN Junying，JIANG Feng，et al. HuatuoGPT，towards taming language model to be a doctor［J］. arXiv preprint arXiv：2305. 15075，2023.
[37] XIONG Honglin，WANG Sheng，ZHU Yitao，et al. DoctorGLM：fine-tuning your Chinese doctor is not a herculean task［J］. arXiv preprint arXiv：2304. 01097，2023.
[38] ZHANG Tianyi，KISHORE V，WU F，et al. Bertscore：evaluating text generation with bert［J］. arXiv preprint arXiv：1904. 09675，2019.

作者简介：王翼虎（ORCID：0009-0004-0501-6770），男，1998 年生，硕士生。研究方向：自然语言处理与人工智能。**白海燕**（ORCID：0000-0002-9552-3845，通信作者，Email：bhy@istic. ac. cn），女，1973 年生，研究馆员。研究方向：信息组织，数字图书馆，关联数据，知识组织系统。**孟旭阳**（ORCID：0000-0003-2853-8008），女，1992 年生，硕士，助理研究员。研究方向：自然语言处理，数字图书馆。

录用日期：2023－06－26