基于意图识别与检索增强生成的校园问答系统

汤博文[1]，马名轩[1]，张以宁[1]，李厚润[1]，温非凡[2]，王达彬[1]，杨加[3]，马皓[3]
（1.北京大学计算机学院，北京 100871；2.北京大学信息科学技术学院，北京 100871；3.北京大学计算中心，北京 100871）

摘　要：为解决传统校园问答系统信息整合能力不足、泛化能力差等问题，设计了基于大语言模型的校园问答系统。使用微调后的大语言模型对用户问题进行意图识别，为不同意图的问题提供有针对性的处理方法，提升用户体验。同时，针对大语言模型生成时的幻觉问题，利用多种校园数据建立了校园知识库，通过检索增强生成方法为模型提供事实依据。实验结果表明，经过指令微调的开源大语言模型可以达到接近甚至超越闭源大语言模型的意图识别准确率。

关键词：大语言模型；检索增强生成；指令微调；意图识别
中图分类号：TP182
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024245

Campus question-answering system based on intent recognition and retrieval-augmented generation

TANG Bowen[1], MA Mingxuan[1], ZHANG Yining[1], LI Hourun[1], WEN Feifan[2], WANG Dabin[1], YANG Jia[3], MA Hao[3]
1. School of Computer Science, Peking University, Beijing 100871, China
2. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
3. Computer Center, Peking University, Beijing 100871, China

Abstract: To address the issues of poor information integration and generalization in traditional campus question-answering systems, a campus question-answering system based on a large language model was designed. The fine-tuned model identified user intents and provided targeted solutions for various types of questions, enhancing the user experience. To mitigate the hallucination problem during language model generation, a knowledge base using diverse campus data was constructed and a retrieval-augmented generation method was employed to ensure factual accuracy. Experimental results indicate that the open-source large language model, after instruction tuning, achieves intent recognition accuracy that is comparable to or even surpasses that of closed-source models.

Keywords: LLM, RAG, supervised fine-tuning, intent recognition

0　引言

在智慧校园建设中，校园智能问答系统是重要的一环。传统问答系统往往以知识图谱、搜索引擎等检索技术作为技术基础[1-2]。这些方法往往面临着语义理解不足和灵活性较差的问题。用户所获得的信息往往分散且相关性不高，还需要进一步进行整合和分析。因此，如何有效地利用召回的知识高效、准确地回答用户问题，成为智能问答系统面临的一大挑战。

近年来，随着人工智能技术的快速发展，尺度定律[3]揭示了大语言模型（LLM, large language model）在进行自然语言处理（NLP, natural language processing）任务时的强大能力，大语言模型在智能对话方面的能力也达到了此前深度学习模型

收稿日期：2024–10–22
通信作者：杨加，yangj@pku.edu.cn

未有的高度。ChatGPT 等新兴智能对话服务均使用大语言模型作为基础[4]。但大语言模型存在着幻觉问题[5]：当没有事实依据作为上下文辅助时，会在对话中生成毫无根据，且真假难分的言论，造成回答质量的下降，甚至对用户产生误导。

为了解决此类问题，研究者们提出了检索增强生成（RAG, retrieval-augmented generation）方法[6]。此类方法同时使用知识库和语言模型，将用户问题在知识库的检索结果交给语言模型作为事实依据，由语言模型对召回结果进行整理和回复的生成。

但检索增强生成方法也有其局限性。大语言模型存在着严格的上下文限制，这就意味着只能将很有限的召回结果提供给大语言模型，因此，检索侧的召回率就成为决定方法性能的关键因素。关键词提取、问题转换、假设文档嵌入[7-8]等方法被用于优化召回率，但这类方法往往带来较高的额外时延，无法在响应速度要求高的实时问答系统中使用。除此之外，检索增强生成方法对多轮对话的处理能力也有一定的缺陷，难以根据历史聊天中的指代信息检索到有效的知识[9-10]。

意图识别也是问答系统中的一项关键技术。意图识别可以将用户的问题进行分类，针对不同的输入意图路由到合理的下一步操作，优化用户体验[11-12]。与此同时，意图路由还可以与检索增强生成结合，一定程度上缩小知识库的范围，降低知识召回和回复生成的难度。文献[13]探索了开源、闭源大语言模型在校园问题意图识别方面的能力，得出了开源模型在仅使用提示词的情况下无法达到 GPT-4 级别的分类准确度。如何提升开源模型的分类能力成为亟待解决的问题。

本文的主要研究工作如下。

1) 设计了一种基于意图识别与检索增强生成的校园问答系统，将检索技术与大语言模型的优势相互结合，可以更加精准、智能地回复用户问题。

2) 利用从校园网收集的文本数据，建立了基于 Elasticsearch[14] 的知识库，结合 BM25（best matching 25）与向量相似度 2 种方式，可以高效检索出有效的知识用于回答生成。同时，提出了基于大语言模型的多轮对话指代消解方法，提升了多轮对话中的检索能力。

3) 提出了一种针对意图识别的大语言模型微调数据构造方法，利用构造的数据集对开源大语言模型进行指令微调，提升了开源大语言模型的意图识别能力。

4) 实验结果表明，在本文的意图识别数据集上，微调后的开源大语言模型在参数量只有闭源模型的十分之一到百分之一的情况下，意图识别准确率可以达到接近甚至超越闭源模型的水平。

1 系统架构

1.1 系统架构与问答流程概述

本系统的整体架构如图 1 所示。智能问答部分由意图识别、知识检索、问题回复 3 个模块组成，而数据处理部分由数据处理模块与基于 Elasticsearch 的知识库组成。当用户提出问题时，首先由意图识别部分将用户的问题进行分类，确定解决问题的路径与知识库；随后在对应的知识库中根据用户问题检索相关知识；最终使用用户问题和相关知识组成提示词，交由大语言模型完成问题回复。

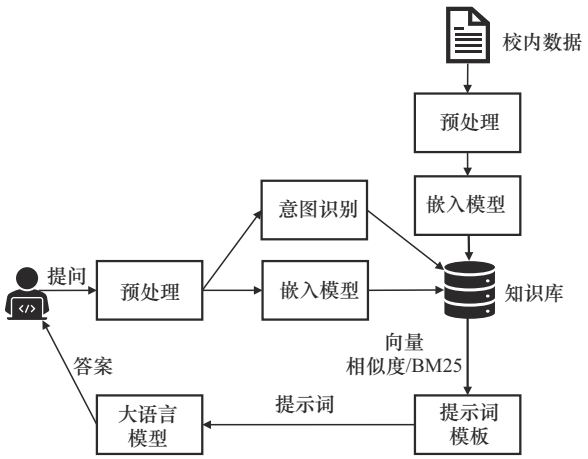


图 1 系统整体架构

1.2 意图识别模块

意图识别模块中，大语言模型通过提示词的辅助，对用户的问题进行分类，从而针对问题类别实施不同的处理方法。本文设计了 2 种不同的处理流程与提示词构造方法来辅助大语言模型完成问题路由的任务。

1) 固定分类法。预设分类集合 S 作为待分类的类别。分别为预设分类提供各自的类别描述作为提示词，让模型根据全部分类描述与用户问题进行分类。

2) 动态分类法。预设分类集合 S 作为待分类的

类别。分类前，首先在预设的完整知识库中召回 top-k 的结果，获得这些结果所属的问题分类，组成分类子集 A，让模型根据 A 中的分类描述与用户问题进行分类。

相比之下，固定分类法减少了一次对知识库的查询操作，同时类别的个数固定，可以使用分类器深度神经网络解决问题。动态分类法可以减少分类类别，从而降低分类难度，同时减少输入的长度，提升推理速度。代价是需要额外进行一次知识库查询，且动态的类别设定无法适用于分类器神经网络，只适用于使用自然语言进行输入输出的语言模型。

考虑到分类任务具有一定的挑战性，且需要大语言模型输出规范的文字以便于后续处理，本文在提示词中使用了链式思维（CoT, chain of thought）[15-16]方法，并加入了一组输入输出实例作为引导。意图识别提示词示例如图 2 所示。

System: 你是北京大学校园问答智能机器人。对于用户的问题，请判断是否可以通过以下表格中的信息回答。表格及简介如下:
{表格信息}
请首先简单写出分析依据，在你输出的最后一行 **只输出一个数字**，作为你认为需要的表项，在这个数字后不要输出任何其他文字。或是**只输出两个字“ 无关”** ，表示单靠一个表格无法完全准确地解决用户的问题。如果你认为多个表格都可以完成需求，也请只选一个输出。
以下是一组问答示例:
[问题]: 学生出差按几类人员报销?
[思考]: 这个问题询问校园网络相关问题，与提供的表格无关。
[答案]: 无关
[问题]: {用户问题}

图 2　意图识别提示词示例

1.3　知识检索模块

知识检索模块中，系统首先根据上一阶段的分类结果选定知识库类别，随后调用大语言模型，利用历史对话对用户问题进行指代消解，生成没有上下文指代信息，适用于知识查找的改写问题。随后，在 Elasticsearch 知识库中检索相应知识，用于后续召回。

召回时，本文使用了 BM25 算法[17]与向量相似度 2 种方法进行召回。BM25 是一种经典的信息检索函数，广泛应用于基于关键字的搜索引擎中，其核心思想是根据查询词的词频、逆文档频率与文档长度计算查询词与文档的相似程度。

基于向量相似度的召回则通过预训练的嵌入模型将文本表示为向量，通过向量间的相似度度量（如余弦距离）判断文本间的相似关系。这类算法能够更好地捕捉文本的语义信息，尤其在词语的多义性和同义性处理上表现优越。

针对检索增强生成方法对多轮对话处理能力的问题，本文设计了指代消解方法，由大语言模型基于历史对话与用户问题生成不含指代的用户问题，提升问题的完整性，从而优化检索效果。

1.4　问题回复模块

问题回复模块中，系统使用知识检索模块召回的知识，与用户问题组合形成提示词，交由大语言模型进行问题回复。问题回复模块使用的大语言模型是百度 ERNIE-4.0，以保证回答的准确性和相关性。

根据意图识别模块分类结果的不同，大语言模型根据知识回答问题的难度也有所不同。大部分分类下，大语言模型只需要根据召回的文本知识进行回复；但在少数复杂问题分类下，模型需要读取表格等非自然语言数据，根据表格信息进行自然语言分析。此时，同样需要使用 1.2 节中提到的链式思维方法，引导大型语言模型先思考后回答，从而使模型提供更清晰、更有说服力的答案，并且帮助用户理解答案背后的逻辑。

1.5　数据处理模块

数据处理模块包括数据的采集、清洗、转换以及最终存入知识库。数据的质量和覆盖范围直接影响到问答系统的性能和准确性。以下是各类数据的情况。

1) 北大网页数据。网页数据主要来源于北大网站群。通过与网站群提供商合作，本文构建了北大网站群数据采集系统，从各网站群提供商提供的接口采集了近 10 年的数据，并定期获取新数据。

2) 网站群之外的网页数据。本文采用爬虫方式，采集了北大新闻网等网站群之外的北大网页数据。

3) 学校相关部门提供的数据。财务部、餐饮中心、保卫部等部门向本文提供了用户咨询时经常会用到的数据。这些数据中，很多是问答类数

据，但也存在非结构化的数据，以及复杂表格数据。

4) 其他文档。包括《本科生手册》《研究生手册》等文档。

对这些数据进行清洗、转换后，本文从语义层面进行二次清洗，去除过时的通知、启事等无用信息。随后对长文本内容按长度进行分块，使用嵌入模型对每个分块进行向量化，存入Elasticsearch中，实现同时支持BM25检索与向量相似度检索。向量化选用的嵌入模型是BGE-M3[18]。这个基于Elasticsearch的知识库覆盖了学校学习、工作、生活的方方面面，不仅支持了问答系统的日常运行，还为后续的模型训练和系统优化提供了坚实的保障。

2 意图识别模型微调与优化

意图识别对于提高问答系统数据检索的质量有重要意义。准确的用户意图识别有助于问答系统更好地理解用户的需求，提供有针对性的解决方案，并进而给出更加准确、有用的回答。

财务数据和餐饮中心提供的数据中，有一些复杂数据，需要特殊的处理。因此，有必要准确识别出用户的提问和查询是否需要使用这些数据进行回答。通过实验发现基于关键词匹配和向量相似度比较的方法不能很好地解决这个问题。因此，本文使用大语言模型作为问答系统的意图识别模块。

在意图识别模块中，系统使用固定分类法和动态分类法对用户的问题意图进行分类。本节尝试针对此任务，对小规模的开源大语言模型进行微调，尝试达到接近甚至超过闭源大语言模型的意图识别能力。

2.1 微调数据集构造

针对意图识别场景，预设了12个问题类别，对应问答系统中需要进行特殊处理的11类问题，以及其他通过检索增强生成方法解决的问题。微调数据集类别名与介绍如表1所示。

根据上述12个类别的信息，本文首先使用GPT-4o模型构造了3 573条高质量问题，每个问题对应到一个类别当中。随后，利用固定分类法和动态分类法的提示词构造方法，分别构造对应的提示词作为大语言模型的输入。最后，使用国内最优的中文闭源大语言模型之一，即百度ERNIE-4.0大语言模型对这些问题进行回答，筛选出分类结果正确的部分，将其输出的链式思维过程以及分类结果一同作为标准回答。微调训练过程中，将让开源模型的输出向对应的标准答案对齐。

2.2 微调方法介绍与微调模型选择

对大语言模型的指令微调使用LLaMA-Factory[19]作为训练框架，使用LoRA[20]方法进行训练。LoRA方法在训练时将预训练模型的模型权重冻结，而使用同规格的低秩矩阵作为可训练权重，在显著降低训练成本的同时保持了有效的学习能力。在微调大语言模型的选取方面，本文选择了Qwen2-1.5B-instruct及Qwen2-7B-instruct的开源大语言模型作为微调的基础模型。相比于参数量级达

表1 微调数据集类别名与介绍

类别名	介绍
日常报销业务报销明细	记录了各个报销项、报销子项，以及每一项的简要说明
国内出差-城市间交通费表	记录了各类人员因公国内出差时，乘坐城市间交通工具（飞机、火车、轮船等）的标准
国际出差-国际交通费表	记录了各类人员因公临时出国时，乘坐国际交通工具（飞机、火车、轮船等）的标准
北京大学国内差旅住宿费限额标准明细表	记录了因公出差时，在国内各个省市的旺季、淡季，各类人员每日差旅住宿费的报销额度
野外考察-城市间交通费表	记录了各类人员因公进行野外考察时，乘坐城市间交通工具（飞机、火车、轮船等）的标准
会议费综合定额开支标准	记录了线上线下的各级别会议的举办开支标准，包括住宿费、伙食费等
校外人员劳务费备案信息	记录了为校外人员发放劳务费时，需要根据人员国籍填写的证件、姓名等信息要求
专家咨询费标准	记录了专家咨询费的发放标准，根据会期、组织形式，发放标准不同
报销开发购置软件费用表	记录了报销开发购置软件费用（包括资产类软件、费用类软件）所需的报销材料
食堂开餐时间表	记录了北京大学每个食堂的每日开餐闭餐时间
暑期各食堂开放时间表	记录了北京大学各个食堂暑期的停伙、开伙时间
无关	上述类别以外的问题

到千亿以上的闭源模型，上述2个模型的参数量只有15亿、70亿，且在同规模模型中有着相对优秀的自然语言处理能力。

3　实验设置与结果

3.1　实验设置

为评估闭源大语言模型，以及微调前后的开源大语言模型在意图识别方面的能力，将2.1节中生成的3 573条数据分为3个类别。

1) 训练集。从ERNIE4.0模型分类正确的问题中，随机选取80%，使用ERNIE4.0模型的回复作为标准回答，用于开源模型微调。

2) 测试集。ERNIE4.0模型分类正确的问题中剩余的20%，用于测试模型的分类能力。

3) 困难测试集。包含ERNIE4.0模型分类错误的问题，作为高难度数据集，同样用于测试模型的分类能力。

本文使用训练集，对所选Qwen2-1.5B-instruct及Qwen2- 7B-instruct开源大语言模型进行LoRA微调，使用的学习率为3×10^{-5}，迭代轮次为1轮。

训练结束后，将微调后的模型分别在测试集、困难测试集以及完整数据集上分别进行分类准确率的测试。

3.2　实验结果与分析

表2和表3分别展示了针对固定分类法和动态分类法进行微调后的模型在2个测试集上的分类准确率。分析表中结果可知，微调后的1.5B与7B模型可以在测试集上达到超过90%的准确率，同时1.5B模型在ERNIE-4.0回答错误的困难数据上也拥有很高的准确率，展示了一定的泛化能力。

表2　微调后模型意图识别准确率——固定分类法

模型（规模）	准确率（测试集）	准确率（困难测试集）
1.5B微调	93.89%	**85.04%**
7B微调	**97.09%**	64.60%

表3　微调后模型意图识别准确率——动态分类法

模型（规模）	准确率（测试集）	准确率（困难测试集）
1.5B微调	97.09%	**85.81%**
7B微调	**97.24%**	44.55%

对比表2和表3可以发现，微调后的1.5B模型使用动态分类法时效果更好，而7B模型使用固定分类法效果相对更好。本文认为，7B模型参数量更大，本身具备强大的自然语言理解能力，动态分类法中分类类别减少带来的优势并不明显；而1.5B模型可塑性更强，其在训练中获得的能力提升更加显著。

为了印证这些猜想，本文在全部数据集上，对微调前后的模型进行了意图识别准确率测试，结果如表4和表5所示。分析结果可知，微调和动态分类均为1.5B和7B的大语言模型带来了分类准确率的提升，但ERNIE4.0使用动态分类法的整体准确率却低于固定分类法。这与之前“模型规模越大，动态分类法中类别减少的优势越小”的猜想相吻合。此外，1.5B模型在微调前的分类准确率远不如7B模型，但微调后却超过了7B模型，这也印证了其训练效果更好的猜想。

表4　各模型意图识别准确率——固定分类法

模型（规模）	准确率（全部数据）
1.5B	34.03%
7B	78.11%
1.5B微调	**94.07%**
7B微调	93.87%
ERNIE4.0	92.33%

表5　各模型意图识别准确率——动态分类法

模型（规模）	准确率（全部数据）
1.5B	41.90%
7B	89.36%
1.5B微调	**96.28%**
7B微调	93.59%
ERNIE4.0	91.49%

结合上述实验可以得出结论：在本文的意图识别数据集上，微调后的Qwen2-1.5B-instruct模型结合动态分类法时，意图识别准确率最高，且模型规模、推理成本均最小，是实际应用中的最佳选择。

3.3　系统实际效果展示

图3展示了系统的实际界面与效果。系统在由大语言模型给出问题回复的同时，也会给出召回的相关网页链接和文档信息，增加回答的可信度。如图3所示，大语言模型可以对多来源信息完成整合，针对用户问题进行准确回答。

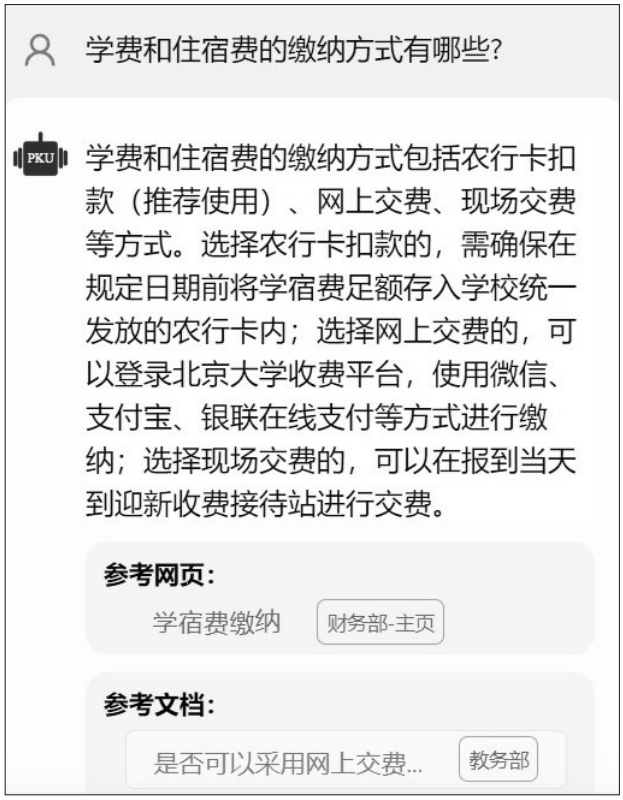


图3 系统的实际界面与效果

4 结束语

本文设计了一种基于意图识别与检索增强生成的智能校园问答系统，将多种校园数据经过清洗和处理后存入知识库，使用意图识别分类用户问题，利用检索增强生成方法缓解大语言模型的幻觉问题，优化生成质量。此外，本文设计了固定分类法与动态分类法2种意图识别方法，构建了校园问答系统意图识别微调数据集，并使用此数据集微调了1.5B与7B规模的开源大语言模型。实验结果表明，在该数据集上，微调后的小规模开源大语言模型在意图识别准确率上达到了接近甚至超越闭源大语言模型的水平。微调后的Qwen2-1.5B-instruct模型结合动态分类法可以达到在实际应用中代替闭源模型的意图识别能力。

参考文献：

[1] 李月, 周江. 一种基于文本相似计算的校园智能问答系统设计[J]. 现代信息科技, 2019, 3(22): 9-12, 17.
LI Y, ZHOU J. Design of a campus intelligent question answering system based on text similarity computing[J]. Modern Information Technology, 2019, 3(22): 9-12, 17.

[2] 龙新征, 郑建宁, 欧阳荣彬, 等. 基于多层策略的校园智能问答系统[J]. 华中科技大学学报(自然科学版), 2016, 44(11): 117-122.
LONG X Z, ZHENG J N, OUYANG R B, et al. University intelligent question answering system based on multi-layer strategy[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2016, 44(11): 117-122.

[3] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[J]. arXiv Preprint, arXiv: 2001.08361v1，2020.

[4] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[J]. arXiv Preprint arXiv: 2303.08774, 2023.

[5] ZHANG Y, LI Y F, CUI L Y, et al. Siren’s song in the AI ocean: a survey on hallucination in large language models[J]. arXiv Preprint, arXiv: 2309.01219v2, 2023.

[6] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[7] ARSLAN M, CRUZ C. Business-RAG: information extraction for business insights[C]//Proceedings of the 21st International Conference on Smart Business Technologies. Setubal: SciTePress, 2024: 88-94.

[8] GAO L Y, MA X G, LIN J, et al. Precise zero-shot dense retrieval without relevance labels[J]. arXiv Preprint, arXiv: 2212.10496v1, 2022.

[9] AGARWAL D, FABBRI A R, RISHER B, et al. Prompt Leakage effect and defense strategies for multi-turn LLM interactions[J]. arXiv Preprint, arXiv: 2404.16251v3, 2024.

[10] CHAN C M, XU C P, YUAN R B, et al. RQ-RAG: learning to refine queries for retrieval augmented generation[J]. arXiv Preprint, arXiv: 2404.00610v1, 2024.

[11] WELD H, HUANG X, LONG S, et al. A survey of joint intent detection and slot-filling models in natural language understanding[J]. arXiv Preprint, arXiv: 2101.08091, 2021.

[12] LOUVAN S, MAGNINI B. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: a survey[J]. arXiv Preprint, arXiv: 2011.00564v1, 2020.

[13] 汤博文, 杨加, 秦辉东, 等. 基于大语言模型的校园网问题解决系统[C]//中国计算机用户协会网络应用分会2023年第二十七届网络新技术与应用年会论文集. 2023: 325-329.
TANG B, YANG J, QIN H, et al. Campus Network Problem-Solving System Based on Large Language Models[C]//Proceedings of the 27th Annual Conference on New Network Technologies and Applications, Network Applications Branch of China Computer Users Association. 2023: 325-329.

[14] ELASTICSEARCH B V. Elasticsearch[J]. Software, Version, 2018, 6 (1) : 1.

[15] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837.

[16] YAO S, ZHAO J, YU D, et al. React: Synergizing reasoning and acting

in language models[J]. arXiv Preprint, arXiv: 2210.03629, 2022.

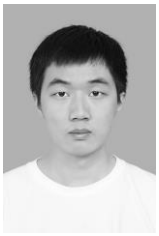
[17] ROBERTSON S E, WALKER S, BEAULIEU M M, et al. Okapi at TREC-4[R]. NIST, 1996.

[18] CHEN J, XIAO S, ZHANG P, ET AL. BGE M3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation[J]. arXiv Preprint, arXiv:2402.03216, 2024.

[19] ZHENG Y, ZHANG R, ZHANG J, et al. Llamafactory: Unified efficient fine-tuning of 100+ language models[J]. arXiv Preprint, arXiv: 2403.13372, 2024.

[20] HU E J, SHEN Y, WALLIS P, et al. Lora: low-rank adaptation of large language models[J]. arXiv Preprint, arXiv:2106.09685, 2021.

[作者简介]



汤博文（2000– ），男，广东梅州人，北京大学硕士生，主要研究方向为人工智能与网络安全。

马名轩（2001– ），男，陕西咸阳人，北京大学硕士生，主要研究方向为人工智能、大语言模型与网络安全。

张以宁（2000– ），男，黑龙江绥化人，北京大学硕士生，主要研究方向为信息检索、自然语言处理、检索增强生成。

李厚润（2001– ），男，重庆人，北京大学硕士生，主要研究方向为计算机应用技术。

温非凡（2002– ），男，安徽肥东人，北京大学本科生，主要研究方向为计算机网络。

王达彬（2002– ），男，湖南邵阳人，北京大学硕士生，主要研究方向为计算机网络。

杨加（1975– ），男，重庆人，博士，北京大学高级工程师，主要研究方向为网络技术应用与数据分析。

马皓（1972– ），男，安徽芜湖人，硕士，北京大学正高级工程师，主要研究方向为计算机网络与信息安全。