● 罗鹏程[1]，王继民[2]，聂 磊[3]
（1. 北京大学图书馆，北京 100871；2. 北京大学信息管理系，北京 100871；3. 北京外国语大学区域与全球治理高等研究院，北京 100089）

基于生成式大语言模型的文献资源自动分类研究*

摘 要：［目的/意义］探索有效提高文献资源自动层次分类和跨语言分类效果的方法。［方法/过程］将文献资源分类视为分类号生成任务，利用图书馆编目数据构造训练集和测试集，基于ChatGLM 3、Llama 2等大语言模型在训练集上进行模型的高效微调，并在中英文测试集上分析模型的分类效果。［结果/结论］在不同的输出格式中，微调大语言模型使其直接输出分类号，可以获得最优的分类效果；随着训练样本数量的增加，微调后的大语言模型分类效果不断提升；基于22000个样本微调的大语言模型在中图法一级类目和完整分类号的准确率分别可达0.8848、0.5076，优于通用大语言模型；在中文文献上训练的大语言模型可以有效地分类英文文献，分类效果仅比中文文献略低；大语言模型生成的分类号中有少量不是有效的中图分类号。

关键词：大语言模型；自动分类；文献资源；层次分类；跨语言分类

DOI：10.16353/j.cnki.1000-7490.2024.12.020

引用格式：罗鹏程，王继民，聂磊．基于生成式大语言模型的文献资源自动分类研究［J］．情报理论与实践，2024，47（12）：174-182.

Research on Automatic Classification of Literature Resources Based on Generative Large Language Model

Luo Pengcheng[1], Wang Jimin[2], Nie Lei[3]
(1. Peking University Library, Beijing 100871; 2. Department of Information Management, Peking University, Beijing 100871; 3. Academy of Regional and Global Governance, Beijing Foreign Studies University, Beijing 100089)

Abstract: [Purpose/significance] Explore effective methods to improve the performance of automatic hierarchical classification and cross-language classification of literature resources. [Method/process] Treat literature resource classification as a classification code generation task, use the library's cataloging data to construct training datasets and test datasets, conduct parameter-efficient fine-tuning of the large language models, such as ChatGLM 3 and Llama 2, on the training dataset, and analyze the classification performance of the model on the Chinese and English test datasets. [Result/conclusion] In different output formats, fine-tuning the large language model to directly output the classification code can obtain the optimal classification performance; as the number of training samples increases, the classification performance of the fine-tuned large language model continues to improve; the accuracy of the fine-tuned large language model based on 22000 samples can reach 0.8848 and 0.5076 respectively for the first-level category and complete classification code of Chinese Library Classification, which is better than the general large language model; the large language models trained on Chinese literature resources can effectively classify English literature resources, and the classification performance is only slightly lower than that of Chinese literature resources. A small number of the classification codes generated by the large language model are not valid Chinese Library Classification Codes.

Keywords: large language model; automatic classification; literature resources; hierarchical classification; cross-language classification

0 引言

对文献资源进行分类组织是图书馆等信息服务机构最为重要的基础性工作。为文献赋予一个合适的分类号需要耗费大量人力，据美国国会图书馆报告显示其十进制分类部门每小时仅能为10.92个书目记录赋予杜威分类号[1]。目前，每年新增的文献资源数量巨大。2022年，我国新出版的图书达到20多万种[2]，中国知网收录的中文期刊论文超过100万篇[3]，此外还有大量学位论文、会议论

* 本文为国家社会科学基金项目“面向多语种社会科学数据的线索发现方法研究”的成果，项目编号：22CTQ025。

文、科技报告、内部资料、进口图书等。面对如此海量的待分类文献资源，仅仅依靠人工分类显得捉襟见肘。

为提高分类标引的效率，文献资源自动分类方法被提出。早期的文本自动分类方法为基于规则的知识工程方法[4]，该类方法在实践中仍有应用，如美国国会图书馆的AutoDewey[5]、上海交通大学图书馆的“自动分类专家系统”[6]。基于规则的方法自适应性差，且规则的编制和维护成本高，目前研究中更多的关注基于机器学习的文献资源自动分类方法，主要采用支持向量机等统计学习模型[7-9]、循环神经网络等深度学习模型[10-11]、BERT 等预训练语言模型[12-14]对文献资源进行扁平分类。考虑到《中国图书馆分类法》（简称《中图法》）等分类体系中的类目具有层次关系，现有研究通常依据分类法自顶向下逐层逐类目构建分类器，将分类器组合以实现文献资源的层次分类[15-17]。这种方法需要训练大量的分类器，并且无法自动确定资源适合的最优层级位置。此外，图情机构通常拥有多语种的文献资源，现有研究中的分类模型不具有跨语言分类能力。本文将文献资源分类视为分类号自动生成任务，通过微调单个生成式大语言模型，理解多语种文献资源的元数据，自动生成最优层级位置的分类号，实现更高质量的文献资源层次分类和跨语言分类。

1 相关研究

文献资源自动分类属于文本自动分类方法在图情领域的应用，已有大量研究对文献资源自动分类方法进行了探索[18]，依据待分类类目之间的关系，可将相关研究分为：扁平分类和层次分类。

1）文献资源扁平分类。文献资源扁平分类是指待分类的类目处于同一层次，类目之间没有上下位关系。早期的相关研究主要利用统计学习模型进行分类，如从《中图法》等分类体系中选取若干同级类目，使用统计学习模型对图书、论文、专利等资源进行分类，从实验结果来看通常支持向量机的分类效果最优[7-9,19-22]。统计学习模型主要基于 TF-IDF 来抽取文本特征，会丢失文本语义信息。深度学习模型可以从训练数据中学习文本嵌入表示，能够保留更多的语义信息，通过利用卷积神经网络、循环神经网络对文献资源进行自动分类，可获得比统计学习模型更优的分类效果[10-11]。深度学习模型需要大量的训练样本，然而实际应用中训练样本的数量常常有限。预训练语言模型能够基于大量文本进行无监督学习，在获得足够好的文本嵌入表示后，再针对特定下游任务进行模型微调。研究者利用 BERT 等预训练语言模型在文献分类数据上进行模型微调，实验结果显示该类模型显著优于统计学习模型、卷积和循环神经网络模型[12-14]。目前文献资源自动分类研究主要关注于扁平分类问题，解决少量平级类目文献资源的自动归类。然而，实际中的《中图法》等分类体系类目数量庞大，且按照层级关系组织类目，扁平分类方法无法实现文献资源按照分类体系的自动组织。

2）文献资源层次分类。文献资源层次分类是指待分类的类目基于分类体系进行组织，类目之间存在上下位层级关系，分类器需要将资源归属到分类体系中某个最合适的类目。在文献资源层次分类中，现有研究依据分类体系自顶向下逐个类目构建分类器，各个分类器组合在一起，将资源从大类开始逐步细化归属到小类[15-16]。该类方法最终会将文献资源归属到最底层的类目，然而在实际应用中，某个资源最合适的类目可能在分类体系的中间层次。Wang[17]提出了一种交互式的分类模型，在每层分类中使用机器学习模型推荐若干概率最大的类目，并由人从其中选择最合适的类目或者停止归类。该方法需要引入人工判断，无法实现完全自动的文献分类。总体来看，目前针对文献资源层次分类的研究较少，现有方法在分类的过程中还无法自动判断资源适合的最优层级类目。

上述研究均是在单一语种上进行模型的训练和测试，而图情机构通常拥有多语种的资源，并且小语种资源的数量较少，不足以构建高精度分类器，现有研究使用的模型缺乏跨语言分类能力。随着 ChatGPT 等大语言模型的出现，其展现出强大的多语言文本理解和生成能力，已有研究探讨了大语言模型在图书馆编目、分类标引等场景的应用前景[23-24]，不过这些研究都主要从宏观层面进行探讨，缺乏系统性的实验分析。

2 研究方法

本文通过收集真实的图书编目数据，以《中图法》为分类依据，基于生成式大语言模型进行分类模型的训练，并与通用大语言模型进行比较分析。图 1 给出了本文的研究框架，主要包括：数据准备、模型训练、模型评价。

1）数据准备。为了对大语言模型进行训练和分类效果评测，本研究从图书馆获取机器可读目录（MAchine Readable Cataloging，MARC）数据，提取图书的标题、摘要、语种、出版年、中图分类号等元数据，基于该数据构造训练集和测试集。为了分析大语言模型不同输出格式对分类效果的影响（如仅输出分类号，或者输出分类号的同时提供对应类目的解释），还需要获取电子版的《中图法》。在训练数据集的构造中，仅选用中文书目数据构造训练样本，同时构造中文测试集。为了直接对比模型在不同语种样本上的分类效果差异情况，本研究将中文测试集中的样本翻译成英文，从而构建英文测试集。通过在中文

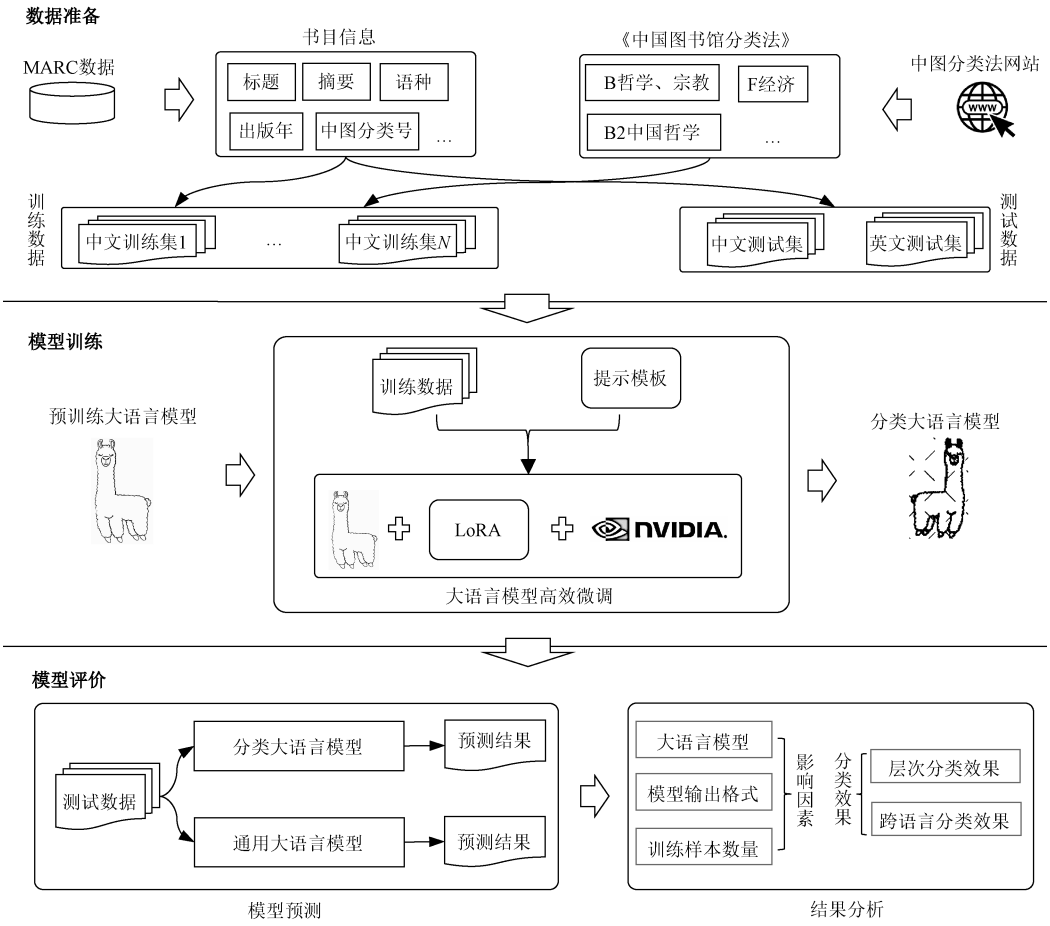

图1 研究框架

Fig. 1 Research framework

训练集上进行模型训练，获得分类大语言模型，并在不同语种的测试集上评测分类效果，以此分析大语言模型的跨语言分类能力。在训练集的构造中，尝试了多种不同类型的分类号输出格式和不同的训练样本量，以便分析不同输出格式和训练样本数量对模型分类效果的影响。

2）模型训练。通用大语言模型可视为通才，具有广泛的知识，但在特定领域的表现可能不专业。《中图法》类目数量巨大，分类规则繁杂，通常需要由专业编目人员才能为文献资源赋予一个合适的分类号。为了使大语言模型在特定领域更专业，本研究利用标注数据微调大语言模型。

大语言模型的训练包括预训练和微调两个阶段。在预训练阶段通常采用自回归变换器（Auto-Regressive Transformer），利用大量无标注的文本数据，进行下一个单词预测训练，使模型获得丰富的知识。在微调阶段，通常先进行有监督微调使得模型学会自然语言问答，然后对模型进行人类反馈的强化学习，使得模型输出与人类价值观保持一致。本文直接选用具有较大影响力的开源多语种预训练大语言模型进行有监督微调，考虑到微调后模型的输出均为分类号相关内容，无须进一步与人类价值观对齐。大语言模型有监督微调本质上是在做文字接龙，需要设计提示模板，以便将输入（书目信息）和输出（分类号）进行组装。参考 Stanford Alpaca[25] 设计了如下提示模板（其中，{{title}} 为文献标题，{{abstract}} 为文献摘要，{{clc}} 为对应文献的分类号输出文本）：

Instruction：

请根据给定文献的元数据信息，为其赋予一个中图分类号。

Input：

【标题】{{title}}

【摘要】{{abstract}}

Response：

{{clc}}

在大语言模型微调中，由于模型参数量巨大，全量微调需要消耗大量的显存和计算资源。为了节省显存和计算开销，本文使用 LoRA[26] 来高效微调模型。神经网络模型通常使用矩阵来表示参数，并通过矩阵运算进行模型计算。LoRA 通过冻结预训练模型的参数，并在少量原始参数矩阵（如注意力层参数矩阵）旁增加可更新参数的矩阵用于模型训练。假设原始参数矩阵 $W_0 \in R^{d\times k}$，LoRA 使用两个矩阵 $B \in R^{d\times r}$ 和 $A \in R^{r\times k}$ 相乘来表示对原始参数矩阵的更新，计算如公式（1）所示：

$$W_0 + \Delta W = W_0 + BA \tag{1}$$

通过将 r 设置为远小于 d 和 k，可以极大减少模型需

要更新的参数量。如假设 d 和 k 为 1024，则参数总量约为 100 万；将 r 设置为 4，则 B 和 A 的参数总量约为 8000。在模型训练初始时，将 A 基于高斯分布随机初始化，B 设置为零矩阵。此时 $\Delta W = 0$，与原始模型完全等价。随后微调时仅对 A 和 B 更新，从而优化模型输出。

3）模型评价。使用微调后的分类大语言模型和通用大语言模型对测试集中的样本进行分类号预测。在预测结果分析中，本文主要采用准确率对模型效果进行评价，计算如公式（2）所示：

$$准确率 = \frac{正确预测的样本数量}{样本总量} \tag{2}$$

此外，在针对各个基本大类的分析中，笔者还采用了精度、召回率和 $F1$ 值进行评价。在分类效果的分析维度上，本文从不同大语言模型、不同类型的输出格式、训练样本数量等因素出发，分析各模型层次分类和跨语言分类效果。

3 实验设置

3.1 训练集与测试集的构造

从图书馆馆藏资源管理系统中获取图书编目数据，通过编写 MARC 解析器，从中提取图书的标题、摘要、中图分类号、语种和出版年等信息。中图分类号可在主表分类号的基础上添加复分号，考虑到主表类目数量众多，与复分号的组合数量巨大，分类难度更高，本文将中图分类号中的复分号部分去除。此外，还从 www. ztflh. com 中采集了电子版的《中图法》，共 45835 个类目。在训练集的构造中，构建了样本数量从小到大的 3 个训练数据集 D_1、D_2、D_3，分别包含 2200、11000、22000 个样本。训练集中的样本均为 2007—2021 年期间出版的中文图书，在样本选取时，按照 22 个基本大类进行随机抽样，D_1、D_2、D_3 中每个基本大类分别有 100、500、1000 个样本。训练集中少量样本有多个分类号，在分类号选取时，本文选择了排在第一位的分类号作为其主要分类号。训练样本的输入为图书的标题和摘要，尝试了 4 种输出格式，如表 1 所示。

表 1　大语言模型训练输出格式示例

Tab. 1　Large language model training output format examples

输出格式	示例
一级类目名	该文献归属于《中图法》的 J 类（艺术）
分类号	该文献分类号为 J06
一级类目名 + 分类号	该文献归属于《中图法》的 J 类（艺术），其分类号为 J06
多级类目名 + 分类号	该文献归属于《中图法》的“J（艺术）> J0（艺术理论）> J06（造型艺术理论）”，其完整分类号为 J06

在测试集的构造中，本文从 2022 年出版的中文图书中随机抽样，构建了包含 660 个样本的中文测试集 T_{zh}，按照 22 个基本大类进行随机抽样，每个基本大类分别有 30 个样本。为了对比分析大语言模型的跨语言分类效果，利用百度翻译 API 将这 660 个样本的标题和摘要翻译成英文，构建了英文测试集 T_{en}。在测试样本分类号选取时，也选择了排在第一位的分类号作为其主要分类号。

3.2 大语言模型的训练

本文直接使用具有较大影响力的开源中英文大语言模型进行有监督微调，包括 ChatGLM 3[27] 和 Llama 2[28]。ChatGLM 3 基于中文和英文进行预训练，能够同时处理中英文自然语言文本。原始的 Llama 2 主要针对英文进行预训练，本研究使用 Colossal-AI 提供的基于 Llama 2 在中文文本上进行增量预训练的 Colossal-LLaMA-2。具体地，在 ChatGLM 3 和 Llama 2 的选择中，使用了参数量在 60 亿 ~ 70 亿量级的预训练模型，即 Huggingface 中 THUDM/chatglm3-6b-base、hpcai-tech/Colossal-LLaMA-2-7b-base。在模型 LoRA 微调时，考虑到标题和摘要均不会太长，将模型最大输入文本长度设置为 768；LoRA 的参数 r 设置为 16；对模型进行 8 比特量化以进一步减少显存需求；训练时的学习率设置为 3×10^{-4}，批次大小为 4，在整个数据集上迭代训练 3 次。

3.3 基于大语言模型的分类号预测

对于微调后的分类大语言模型，将待分类的文献按照微调时所用的模板进行组装（仅空出 {{clc}} 部分），输入大语言模型进行分类号的生成。对于通用大语言模型，本文选择了百度文心一言 4.0（ernie-4.0）和 OpenAI 的 GPT 4（gpt-4-turbo），使用如下提示模板进行分类号预测：

《中国图书馆分类法》是一部在中国广泛使用的、具有代表性的大型综合性分类法。假设您是图书馆的专业编目人员，请您根据图书的标题和摘要，为图书赋予一个中图法分类号。直接给出分类号，不要解释。

```
【标题】{{title}}
【摘要】{{abstract}}
```

在以上模板中，{{title}} 填入文献标题，{{abstract}} 填入文献摘要。由于文心一言总是输出解释文本，使得分类号混杂于其中，因而对于文心一言的调用，还加入了“返回格式为 JSON”的指令，可以极大地减少解释文本的输出。在模型调用中，将 temperature 参数设置为 0（文心一言不能取 0 值，故设置为 0.001），以便获得稳定的输出结果。

4 结果分析

传统分类模型难以用一个模型实现文献资源的层次分类和跨语言分类，本文从这两个方面出发，分析大语言模型的分类效果。

4.1 层次分类效果分析

本节首先基于样本量较少的 D_1 数据集训练模型，分析微调后的大语言模型在一级和多级类目上的分类效果，同时分析不同的大语言模型、不同的模型输出格式对分类效果的影响。然后，选择最优的模型输出格式，基于 D_1、D_2 和 D_3 微调大语言模型，分析训练样本量对模型分类效果的影响。最后，选择 D_3 上微调得到的大语言模型，将其与通用大语言模型进行比较分析，同时分析其在中图法各一级类目下的分类效果。

1）一级类目上的分类效果。图2给出了基于 D_1 微调得到的大语言模型在《中图法》一级类目上的分类准确率。从图2中可以看出，大语言模型分类准确率可达到0.78以上。让大语言模型学习不同类型的输出会影响一级类目的分类准确率，通常输出“一级类目名”“分类号”“一级类目名+分类号”“多级类目名+分类号”的分类效果依次递减。这可能是因为：输出信息越多，大语言模型需要学习的知识越多，而在训练量相同的情况下，模型所学习到的一级类目分类的知识就会相对有所降低。大语言模型直接输出“一级类目名”所需要学习的知识最少，模型在《中图法》一级类目上的分类效果通常最好，ChatGLM 3和Llama 2的准确率分别可达0.8000、0.8182。对比ChatGLM 3和Llama 2可以发现，两个模型的分类效果较为接近，在输出为“一级类目名”时，Llama 2的分类效果较好，而对于其他类型的输出，ChatGLM 3的分类效果则略优。

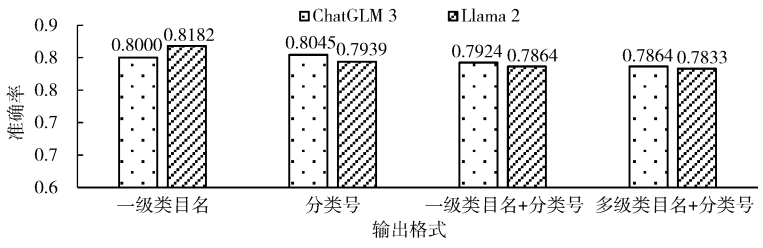

图2 大语言模型在《中图法》一级类目上的分类效果

Fig. 2 Classification performance of large language model on the first-level categories of “Chinese library classification”

2）多级类目上的分类效果。图3给出了基于 D_1 微调得到的大语言模型在《中图法》二级和三级类目上的分类准确率。从图中可以看出，ChatGLM 3和Llama 2在二级类目上的分类准确率最高分别可达0.6167、0.6030，而在三级类目上的准确率最高分别为0.4470、0.4333。从这一结果可以发现，随着分类层级的增加，模型的分类准确率出现了下降。由于 D_1 仅有2200个样本，而《中图法》的二级类目高达200多个，三级类目超过1000个，层级越深，对应的训练样本数量越有限，模型所能学习到的分类知识也越少，导致层级越深分类效果越差。此外，从分类效果来看，ChatGLM 3略好于Llama 2。这可能是因为：ChatGLM 3是原生的中英文大语言模型，在模型预训练中包含了大量中文数据；本文所使用的Llama 2主要基于英文数据进行预训练，并在少量中文数据上做了增量预训练。因而ChatGLM 3的中文理解能力优于Llama 2，故而在中文文献分类上的效果略优。

3）训练样本量对分类效果的影响。从图3可以看出，微调大语言模型使其输出格式为“分类号”通常具有最优的分类效果。因此，后文仅针对输出格式为“分类号”的分类大语言模型进行分析。图4给出了基于 D_1、D_2 和 D_3 微调大语言模型在中文测试集上的分类效果。随着训练样本数量的增加，大语言模型在《中图法》一级、二级和三级类目上的分类效果均呈现增长趋势，其中三级类目的分类效果的提升最为明显。基于 D_3 训练得到的分类大语言模型，最佳的一级、二级、三级类目分类准确率为0.8697、0.7621、0.6621，与 D_1 相比分别提升了约6、15、21个百分点。由此可见，要提升大语言模型的分类效果，需要增加更多的训练样本。此外，对比ChatGLM 3和Llama 2可以发现，当训练样本数量少时，ChatGLM 3的分类效果更优；而当训练样本数量增加时，Llama 2的分类效果更优。

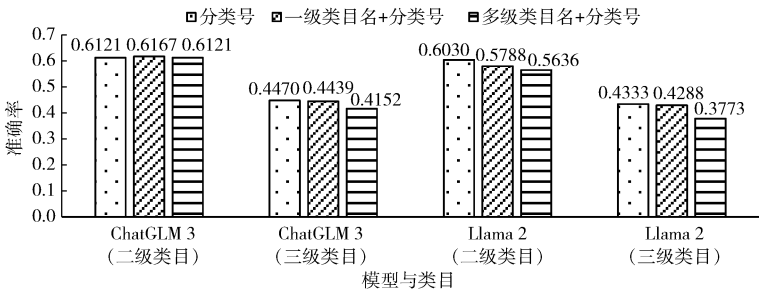

图3 大语言模型在《中图法》二级和三级类目上的分类效果

Fig. 3 Classification performance of large language model on the second-level and third-level categories of “Chinese library classification”

4）与通用大语言模型分类效果的比较分析。图5给出了基于 D_3 训练集微调的大语言模型与通用大语言模型（GPT 4、ERNIE 4.0）分类准确率的比较结

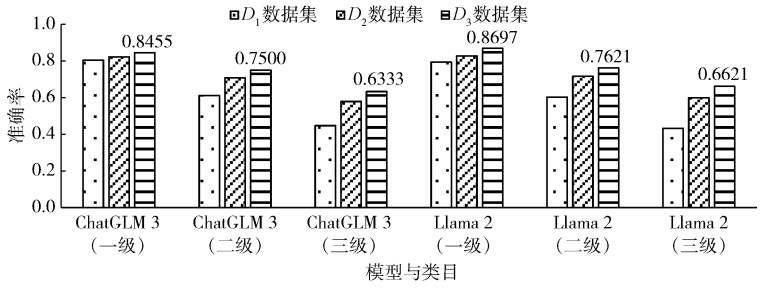

图 4　基于不同训练样本量数据集微调后的大语言模型在中文测试集上的分类效果

Fig. 4　The classification performance of the fine-tuned large language model based on different training sample size datasets on the Chinese test dataset

果。在通用大语言模型中，ERNIE 4.0 在一级、二级、三级类目上分类准确率分别为 0.7152、0.5955、0.5030，完整分类号的准确率也达到 0.2273，均高于 GPT 4。这可能是因为 ERNIE 4.0 预训练中使用了大量中文文献数据，相比而言 GPT 4 预训练中文数据并不占优势，因而模型对中图分类号的理解还不够准确。与 GPT 4 和 ERNIE 4.0 相比，开源的 ChatGLM 3 和 Llama 2 模型规模较小，但是经过微调后，模型的分类能力显著优于通用大语言模型，其中 Llama 2 预测的完整分类号的准确率达到了 0.4076。由此可见，对于文献资源分类任务，通过在领域特定的数据上微调大语言模型，可获得更优的分类效果。

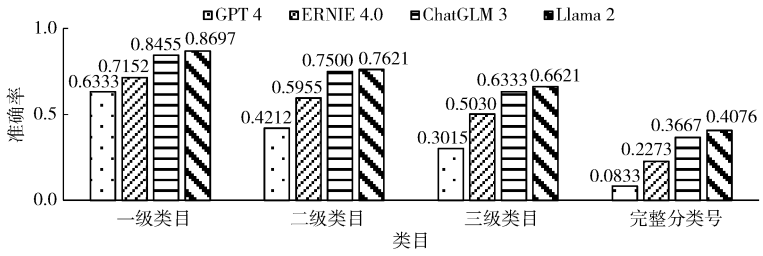

图 5　微调大语言模型与通用大语言模型的分类效果比较

Fig. 5　Comparison of classification performance between fine-tuned large language models and general large language models

5）在各一级类目上的分类效果。不同学科的文献资源特点不同，可能会影响到分类效果。图 6 给出了最优模型（D_3 上微调的 Llama 2 模型）在中文测试集上各一级类目的分类效果。从图中可以看出，不同一级类目的分类效果存在一定的差异。A、U 类的分类效果最好，召回率到达了 1，$F1$ 值最高，均在 0.95 以上。这两个学科的特点较为明显，与其他学科的交叉相对较少，因而分类效果最优。X、E、V、J 类的分类效果也较好，精度、召回率、$F1$ 值均在 0.9 以上。K 类的分类效果最差，但 $F1$ 也达到了 0.79，其他一级类目的 $F1$ 值均在 0.8～0.9 之间。总体来看，各一级类目的分类效果虽然有所差异，但是不存在极端差的情况，最差的 K 类资源的精度、召回率、$F1$ 值达到了 0.8214、0.7667、0.7931，与各指标均值（0.8720、0.8697、0.8686）的差距在 10 个百分点以内。

4.2　跨语言分类效果分析

本文使用的 ChatGLM 3 和 Llama 2 均在中文和英文文本上进行了预训练，能够理解中英文数据。本节分析大语言模型在中文文献

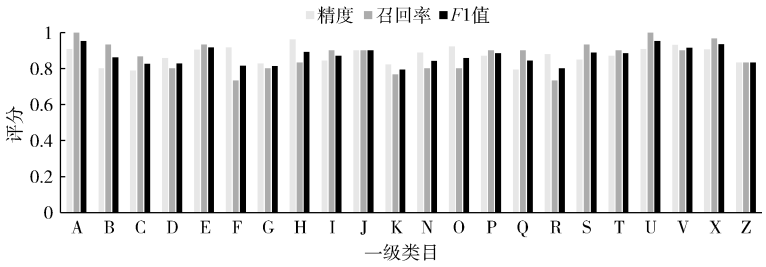

图 6　最优模型在《中图法》各一级类目上的分类效果

Fig. 6　The classification performance of the optimal model on each first-level categories of Chinese library classification

上进行分类训练后，其在英文文献上的分类效果。图 7 给出了基于 D_1 微调得到的大语言模型在英文测试集 T_{en} 上的分类准确率。从图中可以看出，虽然模型是在中文样本上训练得到，但是针对英文样本仍然有较高的准确率。在一级类目上，分类准确率大致在 0.74～0.79 之间，仅比中文分类准确率低大约 2～4 个百分点，相似地在二、三级类目上，英文的分类准确率也仅比中文低大约 2～5 个百分点。此外，从图中也可以看出，针对英文测试样本，输出格式为“分类号”时，大语言模型通常具有最佳的多级类目分类效果。

图 8 给出了 ChatGLM 3 和 Llama 2 基于 D_1、D_2 和 D_3 微调后在英文测试集上的分类效果。从图中可以看出，随着训练样本数量的增加，微调后的大语言模型在英文样本上的分类准确率也在不断提升，Llama 2 在训练样本量较大时比 ChatGLM 3 的分类效果更优。最优模型在英文样本上的一级、二级、三级准确率为 0.8318、0.6924、0.5712，比中文样本的上准确率分别低了约 3、7、9 个百分点。由此可见，即使没有在英文样本上进行模型训练，随着中文样本数量的增加，模型的英文文献分

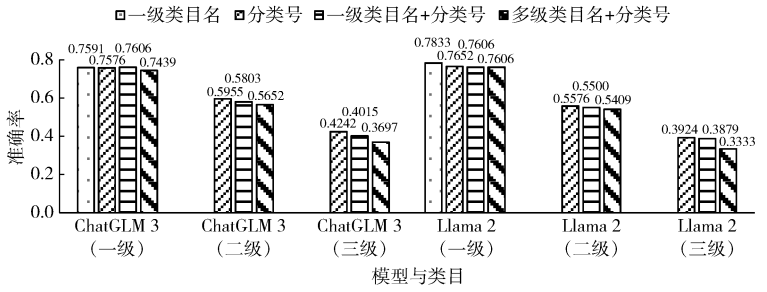

图7 中文样本上训练的大语言模型在英文样本上的分类效果

Fig. 7 The classification performance on English samples of large language model fine-tuned on Chinese samples

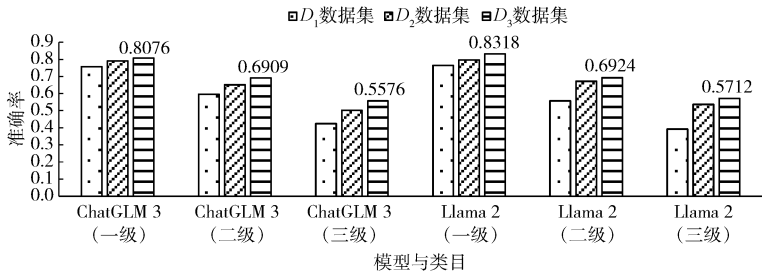

图8 基于不同训练样本量微调后的大语言模型在英文样本上的分类效果

Fig. 8 The Classification performance of large language model on English samples after fine-tuning based on datasets with different training sample sizes

类能力也在不断提升。

从前面关于大语言模型跨语言分类效果的分析可以发现，通过微调多语种大语言模型，即使没有特定语种的训练样本，模型也能够以相对较高的准确率来预测该语种文献资源的分类号，同时预测效果也随着其他语种训练样本的增多而提高。在实践应用中，可以利用多语种大语言模型在中文、英文以及其他小语种资源上进行模型训练，所得到的分类大语言模型能够在所有语种上取得较好的分类效果。

4.3 讨论

从以上分析可以发现，基于生成式大语言模型进行微调，可以实现更为准确的文献资源层次分类和跨语言分类，下面从分类号的有效性、分类号的准确性和分类模型的应用三方面进行讨论。

1）分类号的有效性。传统分类模型在训练和预测时的类别取值范围确定，然而大语言模型理论上可以生成任意文本。本研究将最优模型（基于 D_3 微调的 Llama 2）在测试集 T_{zh} 上预测的分类号与训练集 D_3 中的分类号进行比较分析，结果发现在预测的476个不同的中图分类号中，有36个（占比7.56%）分类号通过开头匹配（如预测分类号为R9，训练集中存在R917，则匹配成功）没有出现在训练集中。这反映出大语言模型具有一定的创造性，由于《中图法》类目数量众多，训练集中很难涵盖所有可能的分类号，大语言模型的这种创造性具有应用价值。然而目前其准确率较低，在新分类号中，仅有两个样例预测正确。部分新分类号可能不是有效的中图分类号，分析发现有47.22%的新分类号无效，如K152.07、R999.1等。在这些无效的分类号中，前面若干级类目的取值正确，后面部分取值无效，如R999.1中的“R99”有效，后面的“9.1”无效。图9给出了新分类号不同类目深度对应的通过开头匹配在训练集中出现的新分类号数量。可以看出在这36个新分类号中，一级类目和二级类目均在训练集中出现过，随着类目层级深度的增加，越来越多的分类号没有出现在训练集中。这可能是因为：训练集中包含一二级类目的样本量较多，模型以较大的概率学习生成正确的一二级类目，随着类目层级深度的增加，相应类目对应的样本越来越少，模型生成未见到过类目的概率逐渐增加。在实际应用中，可以考虑引入《中图法》作为知识库，对生成的分类号的有效性进行验证，并通过截断无效部分来确保分类号的正确性。

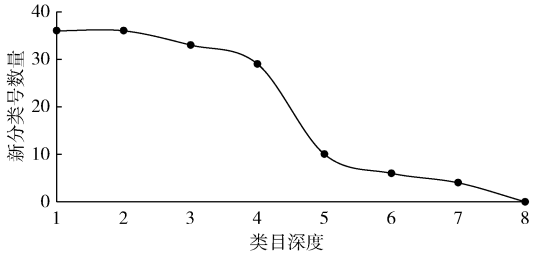

图9 通过开头匹配在训练集中出现的新分类号数量

Fig. 9 The number of new classification codes that appear in the training set by starting matching

2）分类号的准确性。大语言模型能够生成完整的分类号，本文最优模型在测试集 T_{zh} 上预测分类号的准确率可达到0.4076，相应的中图法一级、二级、三级类目分类准确率为0.8697、0.7621、0.6621。测试集中少量样本有多个分类号，在分类效果评价中仅选用了排名第一位的分类号，如果将所有分类号进行考虑，则完整分类号的准确率可达到0.4227，相应的中图法一级、二级、三级类目分类准确率分别为0.8848、0.7848、0.6894。考虑到大语言模型生成的分类号可能没有做到细分，只是真实分类号的前一部分。本研究将预测值与所有真实分类号进行开头匹配，其准确率为0.5076。由此可见，微调后的大语言模型

所生成的分类号中约有一半预测正确。

一本图书可能描述了多个主题，与分类体系中多个类目相关，不同人对图书内容的理解不同，很可能为相同的图书赋予不同的分类号。为此，对最优模型在测试集 T_{zh} 上一级类目分类错误的 76 个样例进行分析，对比它们在国家图书馆、上海图书馆、广州图书馆馆藏目录中的分类号。结果发现，有 5 个样例的预测分类号与公共图书馆的分类号一致，另有两个样例的一级类目一致。对于其余的样例，发现模型预测分类号与图书内容通常也有较大的相关性。如表 2 所示，样例 1 ~ 5 的预测分类号 A81（马克思主义的学习和研究）、Q4（生理学）、TV213.4（水利资源的管理、保护与改造）、C934（决策学）、D929（中国法制史）与图书的主题内容相关性较大。由此可见，所谓分类错误的样例，其分类号在某种程度上也具有一定的正确性。

表 2　分类号预测错误的样例

Tab. 2　Examples with incorrectly predicted classification code

#	图书题名	真实值	预测值
1	马克思主义学术中国化研究	D61	A81
2	生理学	R33	Q4
3	人民胜利渠志	S279.261	TV213.4
4	基于犹豫模糊信息的综合评价理论、方法与应用	O159	C934
5	民国璧山司法档案案卷提要，民事卷	Z89	D929

3）分类模型的应用。为了提高模型的应用价值，需要进一步提升模型的分类准确率。从前面的实验分析可知，通过增加更多的训练样本，模型可以实现更优的分类效果。由于大语言模型的微调非常耗时，本文仅在 22000 个样本上进行训练，考虑到大型图书馆通常拥有数百万的编目数据，利用全量数据进行模型训练可极大提升分类准确率。本文所用大语言模型的参数量在 70 亿左右，可在消费级显卡上运行，能够以较低的成本支撑图情机构文献资源的自动分类。本文提出的分类大语言模型可为文献资源生成分类号，辅助馆员进行分类标引，并支撑图书馆相关服务。通过在编目系统中嵌入模型，能够智能地为编目员推荐分类号。考虑到本研究中的分类大语言模型生成完整分类号的准确率可达 0.4227（开头匹配准确率 0.5076），因而理论上至少有一半的分类号可被编目员接受。在馆藏资源相关统计中，常常需要使用中图分类号来识别图书的学科。存在大量没有中图分类号的馆藏资源（如老旧资源）会使得统计不准确，利用分类大语言模型为其生成分类号，可在一定程度上优化馆藏资源的统计与分析。在馆藏资源检索中，常需要提供学科筛选功能，大语言模型生成的分类号能够支撑大量缺乏分类号的馆藏资源进行筛选，帮助用户更高效的获取文献。

5　结论与展望

为了实现文献资源的层次分类和跨语种分类，本文将文献分类视为分类号生成任务，利用大语言模型为文献资源生成分类号。通过基于图书馆真实的文献资源编目数据构造训练集和测试集，利用 ChatGLM 3、Llama 2 在训练集上进行模型的高效微调，并在中英文测试集上分析模型的分类效果。实验结果表明：在不同的输出格式中，微调大语言模型使其直接输出分类号，可以获得最优的分类效果；与通用大语言模型相比，微调后的大语言模型分类效果更优；随着训练样本数量的增加，大语言模型的分类效果不断提升；基于 22000 个样本微调得到的分类大语言模型在中图法一级、二级、三级类目上的分类准确率可达 0.8848、0.7848、0.6894，完整分类号通过开头匹配的准确率为 0.5076；不同学科文献的分类效果存在一定差异，其中 A、U 类的分类效果最好；在中文样本上训练得到的分类大语言模型可以有效地分类英文文献，其分类效果仅比中文文献略低，具有较好的跨语言分类能力。本文仅在数量有限的训练样本上针对中英文大语言模型进行了模型高效微调和分类效果分析，验证了大语言模型在文献资源的层次分类和跨语种分类上的优势。为了实现更好的分类效果，未来可在更大规模（如百万量级）的训练样本上微调多语种（包含中、英、日、韩、德等）大语言模型。此外，本文在大语言模型的分类应用中没有利用外部知识库，考虑到专业编目人员在编目时会参考《中图法》，未来可将《中图法》作为外部知识库输入大语言模型，来提升模型分类效果，同时解决生成无效分类号的问题。□

参考文献

[1] Library of Congress. Acquisitions and bibliographic access directorate [R/OL]. [2024-05-10]. https://www.loc.gov/catdir/aba07.pdf.

[2] 国家统计局. 中国统计年鉴 [M/OL]. [2024-05-10]. http://www.stats.gov.cn/sj/ndsj/2023/indexch.htm. (National Bureau of Statistics of China. China statistical yearbook [M/OL]. [2024-05-10]. http://www.stats.gov.cn/sj/ndsj/2023/indexch.htm.)

[3] 中国知网 [DB/OL]. [2024-05-10]. https://www.cnki.net/. (China national knowledge infrastructure [DB/OL]. [2024-05-10]. https://www.cnki.net/.)

[4] SEBASTIANI F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34 (1): 1-47.

[5] Library of Congress. AutoDewey [EB/OL]. [2024-05-10]. https://www.loc.gov/aba/dewey/practices/autodewey.html.

[6] 上海交通大学图书馆. 自动映射分类专家系统 [EB/OL]. [2024-05-10]. http://clc.lib.sjtu.edu.cn/. (Shanghai Jiao

Tong University Library. Automatic mapping and classification expert system [EB/OL]. [2024-05-10]. http://clc.lib.sjtu.edu.cn/.)

[7] KRAGELJ M, BORŠTNAR K M. Automatic classification of older electronic texts into the universal decimal classification-UDC [J]. Journal of Documentation, 2021, 77 (3): 755-776.

[8] GOLUB K, HAGELBÄCK J, ARDÖ A. Automatic classification of Swedish metadata using dewey decimal classification: a comparison of approaches [J]. Journal of Data and Information Science, 2020, 5 (1): 18-38.

[9] CASSIDY C. Parameter tuning Naïve Bayes for automatic patent classification [J]. World Patent Information, 2020, 61: 1-7.

[10] LYU Lucheng, HAN Tao. A comparative study of Chinese patent literature automatic classification based on deep learning [C] //2019 ACM/IEEE Joint Conference on Digital Libraries. New York: ACM, 2019: 345-346.

[11] GIANNOPOULOU E, MITROU N. An AI-based methodology for the automatic classification of a multiclass ebook collection using information from the tables of contents [J]. IEEE Access, 2020, 8: 218658-218675.

[12] 罗鹏程, 王一博, 王继民. 基于深度预训练语言模型的文献学科自动分类研究 [J]. 情报学报, 2020, 39 (10): 1046-1059. (LUO Pengcheng, WANG Yibo, WANG Jimin. Automatic discipline classification for scientific papers based on a deep pre-training language model [J]. Journal of the China Society for Scientific and Technical Information, 2020, 39 (10): 1046-1059.)

[13] 蒋彦廷. 依据《中国图书馆分类法》的英文图书分类探索 [J]. 北京大学学报 (自然科学版), 2023, 59 (1): 11-20. (JIANG Yanting. English books automatic classification according to CLC [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2023, 59 (1): 11-20.)

[14] ABDELRAHMAN E, FOX E. Improving accessibility to Arabic etds using automatic classification [C] //International Conference on Theory and Practice of Digital Libraries. Cham: Springer International Publishing, 2022: 230-242.

[15] 张智雄, 赵旸, 刘欢. 构建面向实际应用的科技文献自动分类引擎 [J]. 中国图书馆学报, 2022, 48 (4): 104-115. (ZHANG Zhixiong, ZHAO Yang, LIU Huan. Construction of a practical application-oriented automatic classification engine for scientific literature [J]. Journal of Library Science in China, 2022, 48 (4): 104-115.)

[16] GOMEZ J C. Analysis of the effect of data properties in automated patent classification [J]. Scientometrics, 2019, 121 (3): 1239-1268.

[17] WANG Jun. An extensive study on automated Dewey decimal classification [J]. Journal of the American Society for Information Science and Technology, 2009, 60 (11), 2269-2286.

[18] DESALE K S, KUMBHAR M R. Research on automatic classification of documents in library environment: a literature review [J]. Knowledge Organization, 2013, 40 (5): 295-304.

[19] 王昊, 严明, 苏新宁. 基于机器学习的中文书目自动分类研究 [J]. 中国图书馆学报, 2010, 36 (6): 28-39. (WANG Hao, YAN Ming, SU Xinning. Research on automatic classification of Chinese bibliographies based on machine learning [J]. Journal of Library Science in China, 2010, 36 (6): 28-39.)

[20] 杨敏, 谷俊. 基于SVM的中文书目自动分类及应用研究 [J]. 图书情报工作, 2012, 56 (9): 114-119. (YANG Min, GU Jun. Study and apply of Chinese bibliographies automatic classification based on support vector machine [J]. Library and Information Service, 2012, 56 (9): 114-119.)

[21] YU HUAFENG. Bibliographic automatic classification algorithm based on semantic space transformation [J]. Multimedia Tools and Applications, 2020, 79 (3): 9283-9297.

[22] WU Mingfang, LIU Y H, BROWNLEE R, et al. Evaluating utility and automatic classification of subject metadata from research data Australia [J]. Knowledge Organization, 2021, 48 (3): 219-230.

[23] LUND D B, WANG Ting. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? [J]. Library Hi Tech News, 2023, 40 (3): 26-29.

[24] 曹茹烨, 曹树金. ChatGPT完成知识组织任务的效果及启示 [J]. 情报资料工作, 2023, 44 (5): 18-27. (CAO Ruye, CAO Shujin. The effect and enlightenment of ChatGPT in completing knowledge organization tasks [J]. Information and Documentation Services, 2023, 44 (5): 18-27.)

[25] Stanford alpaca: an instruction-following LLaMA model [EB/OL]. [2024-05-10]. https://github.com/tatsu-lab/stanford_alpaca.

[26] HU E J, SHEN Yelong, WALLIS P, et al. Lora: low-rank adaptation of large language models [R/OL]. (2021-10-16) [2024-05-10]. https://arxiv.org/abs/2106.09685.

[27] ChatGLM3 series: open bilingual chat LLMs [CP/OL]. [2024-05-10]. https://github.com/THUDM/ChatGLM3/.

[28] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundation and fine-tuned chat models [R/OL]. (2023-07-19) [2024-05-10]. https://arxiv.org/abs/2307.09288.

作者简介：罗鹏程 (ORCID: 0000-0001-9598-0715), 男, 1989年生, 博士, 副研究馆员。研究方向: 学术数据挖掘, 科学数据管理, 开放获取等。**王继民** (ORCID: 0000-0002-3573-7788), 男, 1966年生, 博士, 教授, 博士生导师。研究方向: 机器学习, Web挖掘, 科学评价等。**聂磊** (ORCID: 0000-0003-1995-4114, 通信作者, Email: nielei@bfsu.edu.cn), 男, 1989年生, 博士, 讲师。研究方向: 数据管理与应用。

作者贡献声明：罗鹏程, 设计研究方案, 实验设计与结果分析, 论文撰写。**王继民**, 提出研究思路, 论文修改。**聂磊**, 实验设计, 论文修改。

录用日期： 2024-09-03