面向图书自动分类的大语言模型提示学习研究*

戎 璐

【摘 要】为避免高昂的人力成本，从少量样本中学习图书类目的特征与规律已成为图书馆学的热门问题。以图书为研究对象，利用大语言模型 ChatGPT 表征文本，构建大语言模型提示学习模型，以实现自动识别并分类图书的目的。针对当前方法需要大规模数据样本与大量训练时间的缺陷，构建“询问大语言模型 – 提示 – 生成”图书分类范式。在广州图书馆和郑州图书馆 10 个一级类目共 114 823 条图书数据集上进行实验验证。实验结果显示，此范式在精准率、召回率与 F1 等指标上获得最优分类结果。

【关键词】提示学习 图书分类 小样本学习 图书资源

Abstract：In order to avoid high costs of manpower，learning the features and patterns of book categories from a small number of sampleshas become a brand – new problem in library science. This paper takes the books as our research target，uses the large language model（ChatGPT）to represent the texts and images，builds a large language model based prompt learning model for recognizing and classifying books. In view of the defects of the current method requiring large – scale data samples and training time，this paper proposes a new paradigm of“inquiry large language model – prompt – generation”. The model was verified experimentally on the dataset of 114 823 books from Guangzhou Library and Zhengzhou Library，and the best results were obtained in terms of precision，recall and F1.

Key words：prompt learning book classification few – shot learning book resource

DOI:10.15941/j.cnki.issn1001-0424.2024.01.012

0 引言

据国家统计局最新国民经济公报显示，2021 年全国出版各类报纸 276 亿份，各类期刊 20 亿册，图书 110 亿册，同比增长约 8. 9%。随着国民经济增长、文化消费支出增加以及国家产业政策的有力引导，图书市场规模呈现蓬勃发展的态势，主要体现在图书数量的日趋增长与图书种类的丰富多样。因此，高效率组织与分类已经是图书信息管理、数字化建设等领域的热点课题，是构建图书馆各项业务的基础，同时也是一项具有挑战性的任务[1]。

传统人工标注的方式已经暴露出诸多局限[2]，主要是：（1）标注准确率受人专业经验的限制。人工标注方法需要图书分类员具备深厚的专业知识和经验，以便能够对不同书籍进行准确分类。但是，专业人员的经验积累需要很长时间，而且很难量化和标准化，这种方法的可扩展性和可重复性较差；（2）人力成本高。人工标注方法需要大量人力，特别是对于大规模图书分类任务而言，需要大量图书馆学专业人员完成；（3）时间成本高。人工标注方法需要大量时间，特别是难以满足紧急情况或需要快速分类的场景。

如今，图书自动分类研究应运而生。图书自动分类旨在提高图书分类效率与准确性，满足大规模图书组织与分类需求。图书自动分类使用自然语言处理技术，对图书题目、摘要、目录等信息进行分析和处理，将图书分门别类组织，不仅可以提高分类效率，还可以减少人为错误和主观因素的影响，提高分类一致性和标准化。因此，图书自动分类受到学术界与工业界的广泛关注。

发展至今，因机器学习算法过度依赖前期特征工程且分类准确率较低等原因，人们对其的关注度越来越少。而深度学习算法凭借端到端式架构优势，已经成为图书自动分类领域的主流范式。深度学习算法具有优良

* 本文系国家自然科学基金青年基金项目“面向互动对话的类量子情感分析模型研究”（项目编号：62006212）、河南省重点研发与推广专项项目“量子概率驱动的多模态多任务情感识别模型”（项目编号：222102210031）的研究成果之一。

的表达能力，可以自动从图书数据中学习到复杂的特征和模式，在各种类型的任务上获得优良性能。然而，已有的基于深度学习的方法也存在若干缺陷[3]：（1）数据要求高。深度学习模型需要大量的标注数据进行训练，否则容易出现过拟合现象；（2）训练成本高。深度学习模型需要在强大的计算资源上进行训练，训练时间长，对时间敏感性场景不够友好；（3）参数寻优难度大。深度学习模型内置大规模超参数需要寻优（例如 GPT－4.0 拥有 1.8 万亿参数），只有经验丰富的专家才能够进行有效的调试，对非专业人士不友好。此外，深度学习模型对硬件设备要求非常严苛，GPT－4.0 训练一次需要 25 000 块 GPU 计算单元，成本约为 6 500 万美元；（4）可理解性差。深度学习模型架构往往比较复杂，难以解释其内部运作的原理和逻辑，图书馆员往往难以理解其机制。

随着 GPU 算力的不断提升，深度学习模型已经开始迈入大语言模型（large language model，LLM）时代。当模型规模较小时，模型的性能和参数符合比例定律，即模型的性能提升和参数增长基本呈线性关系。然而，当模型的参数量一旦突破千亿、万亿级别（例如 GPT－3、ChatGPT、GPT－4、GPT－4 Turbo），其语言与推理能力将获得质的飞跃，这些能力被称为大语言模型的“涌现能力”（如理解人类指令、提示生成能力等）。2022 年底，大语言模型 ChatGPT 的问世，一举改变了传统研究者的模型使用方式，成为人工智能新的里程碑，仅仅经过两个月，月活跃用户已达到 1 亿，是史上用户增速最快的人工智能应用。2023 年 11 月，OpenAI 正式发布 GPT－4 Turbo，成为目前最强大的通用人工智能助理。

传统的深度学习方法主要通过在大规模下游任务数据上重新训练或微调以获得结果。而大语言模型存在诸多研究难题，例如参数规模过于庞大，下游任务难以收集足够训练数据，且受限于硬件设备，使得参数更新非常困难。鉴于现有图书自动分类技术面临的挑战，本文提出一个关键研究问题：如何有效利用大语言模型，尤其是基于提示学习（prompt learning）的方法，克服现在深度学习在图书分类方面的局限性？该问题的核心在于探索一种新的方法，它既能保持分类的高准确性，又能降低对庞大标注数据集的依赖，减少训练时间和成本。这一问题的探讨对于实现更加高效和用户友好的图书自动分类系统至关重要，对于图书馆的智能化建设具有重要意义。基于大语言模型的提示学习范式提供了一个全新的视角，它具有解决现有方法诸多挑战的潜力，为图书自动分类提供一种更加高效、经济且易于操作的解决方案。

本研究的主要目的是探索基于大语言模型的提示学习范式在图书自动分类中的有效性和实用性，旨在构建一种图书自动分类新方法，即从现有的大规模图书样本－训练模型－更新参数－分类转变为小样本－询问大语言模型－提示－生成类目。该方法结合了大语言模型的强大语言处理能力和提示学习的灵活性，以期解决传统深度学习方法中的数据依赖、高成本和低可解释性问题。该方法具有以下优点：（1）图书分类将摆脱复杂的模型训练与参数调节；（2）无需生成大规模训练图书样本集；（3）不显著性改变大模型内部结构；（4）短时间内即可完成图书分类。两种范式对比如图 1 所示。通过采用基于大语言模型的提示学习，我们可以在保持高准确性的同时降低对大规模标注数据的依赖，减少计算资源的需求，提高模型的可解释性和用户友好性。此外，该方法的成功实施将为其他领域提供借鉴，展示了大语言模型在处理复杂信息分类问题中的潜力，从而推动人工智能技术在更广泛领域的应用和发展。

具体而言，本研究提出的基于大语言模型提示学习的图书自动分类方案，可以实施两种执行思路：（1）通过预训练语言模型 BERT 与残差神经网络（Residual Neural Network，ResNet）提取图书的两种模态特征，即文本（例如标题、关键词与简介）与图像（封面图）特征，设计 5 种典型提示模板，采用注意力机制融合以上提示模板。将该模板直接输入预训练语言模型 RoBERTa，确定图书类目，达到自动图书分类的目的；（2）设计 5 种典型提示模板，直接调用基于 GPT－3.5 的 ChatGPT API 接口，将提示模板作为用户提问输入，ChatGPT 将生成 5 种图书类目答案，根据最大投票机制选择最终图书类目。

本研究根据《中图法》第五版分类体系[2]，在广州图书馆和郑州图书馆分别收集了 114 823 本与 200 本图书作为一大一小两个数据集。大规模数据集旨在验证提示学习在大规模图书自动分类任务上的有效性，而小样本数据集旨在验证提示学习可以避免大规模模型训练，而直接获得图书分类结果。本研究根据研究方案与其他机器学习、深度学习方法比较，验证其真实效果，探索基于提示学习的图书自动化分类的潜力。值得说明的是，本研究关注小样本图书分类的原因并不是因为提示学习只能应用于小规模图书分类场景，而是为了展示提示学习与传统深度学习不同。提示学习不再要求图书馆员进行模型训练，不强制图书馆员收集大规模训练数据集，能摆脱传统深度学习方法的限制，可应用于大规模图书分类场景。

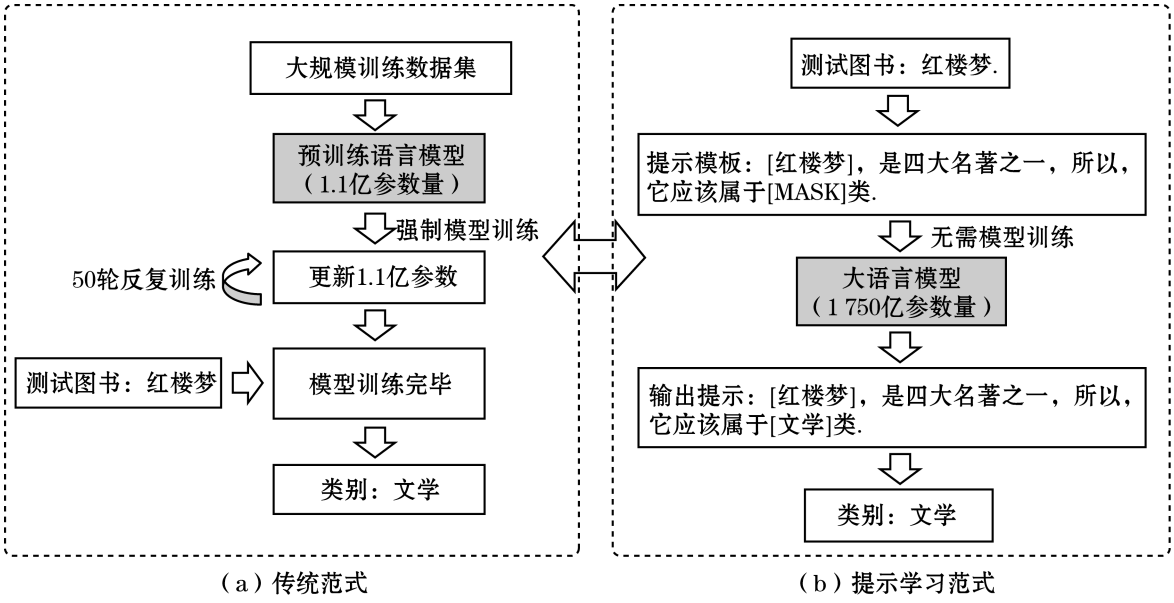


图 1　现有范式与提示学习范式对比

1　相关研究现状

为方便理解后续内容，本文先解释相关专业术语，其次简述机器学习与深度学习的图书自动分类相关研究方案。

1.1　相关专业术语

（1）提示学习：指利用一定的自然语言提示或模板来指导模型进行预测的方法。在提示学习中，模型接收一些提示或模板，然后根据这些提示生成相关的文本或完成相关的任务，该方法的优点是可以通过提示的方式引导模型生成更加合理、准确的文本。提示学习在最近几年中得到了广泛的应用，例如在 GPT－3 等自然语言处理模型中就广泛使用了这种方法，取得了很好的效果[4]。

（2）ChatGPT：是由 OpenAI 公司开发和训练的一种基于 GPT－4. 0 架构的大型语言模型（2023 年 11 月 6 日已经公布 GPT4. 0－Turbo 为基座的 ChatGPT，这也是截至本研究结束时最新的版本，是一个由 GPT－4 升级的全新语言模型），使用了深度神经网络进行训练。模型结构是 decoder－only 的 Transformer 语言模型。生成方式是基于条件生成（conditional generation）的思想，即在给定一些输入文本的情况下，预测并生成下一句话。生成的过程是通过对输入文本进行编码，然后解码生成下一句话的过程。在生成过程中，模型会考虑到前文的语境，根据前文中出现的单词和短语的频率和位置等信息，来预测下一句话的内容。它运行在不同的平台上，包括网站、应用程序、聊天机器人等，其目标是为用户提供高质量、个性化的对话体验，让人们可以更轻松地获取所需的信息、解决问题、进行交流和娱乐。以 ChatGPT、GPT 4－Turbo 为代表的大语言模型目前已经成为人工智能领域最前沿的研究课题，吸引了来自学术界与工业界的全部目光。笔者致力于将其应用于图书分类场景，剖析其性能与效率。

1.2　基于机器学习的图书分类

20 世纪 80 年代以来，统计学习与机器学习逐渐兴起，其技术如支持向量机（SVM）、随机森林（RF）及朴素贝叶斯（NB）等，普遍应用于图书馆学领域。例如：张坤等[5]详细讨论了机器学习算法在图书馆学领域的发展脉络与应用现状，并分析其未来发展趋势。刘高军等[6]通过主题模型提取图书的主题与摘要信息作为混合特征，输入极限学习机中进行图书自动分类，旨在提高中文图书自动分类的效率。赵萌等[7]提出一种基于增量学习的图书文本分类方法，该方法利用现有的图书资源创建增量学习模型，从而能够持续不断地帮助模型学习新的图书类目。Gu Shiqi 等[8]探讨梯度下降树算法在图书分类中的地位，通过与其他机器学习对比，获得最优性能。Safae L 等[9]在网页文本分类任务上展开实验，从实证角度对比三种常用的机器学习算法。在他们的数据集上，支持向量机成为最佳分类器。Shah K 等[10]设计出一个新闻事件类目检测算法。在分类器实现部分，作者分别选择并比较了逻辑回归、随机森林和 K 最近邻分类器。

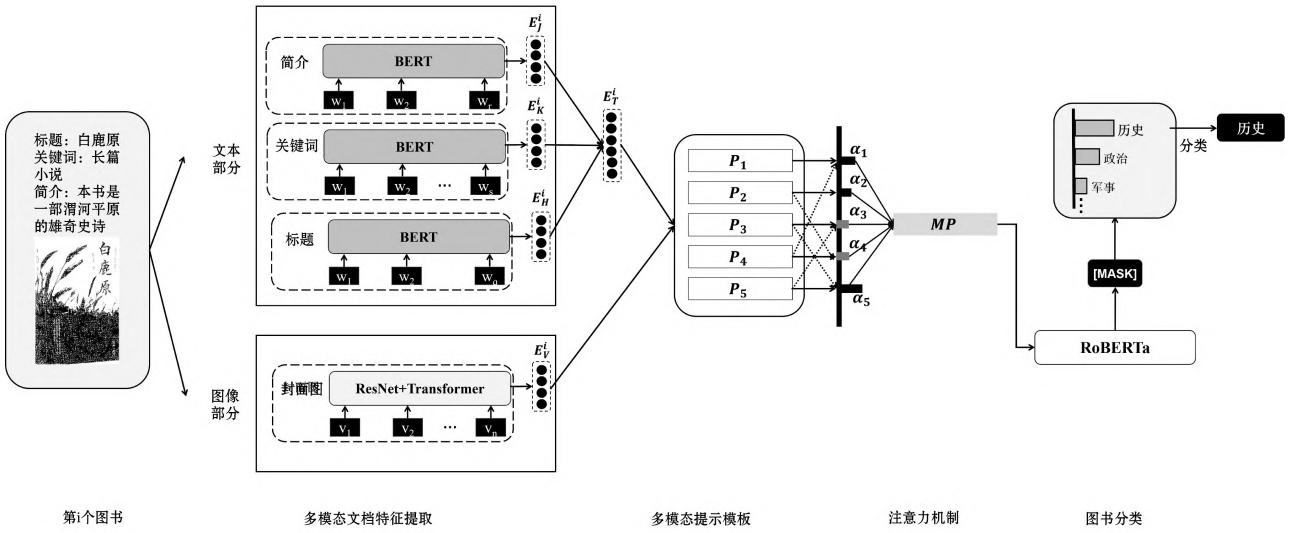
总而言之，基于机器学习的方法可以从数据中学习近似规律，但该方法严重依赖于特征工程与分类器的效果。

1.3　基于深度学习的图书分类

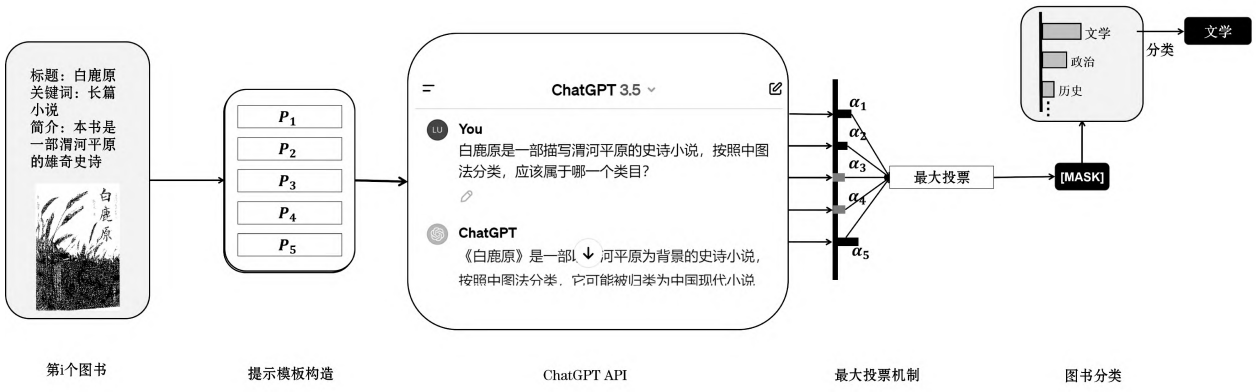
凭借逐层抽象与端到端结构优势，深度学习通常能够获得更高的分类准确率。Kim Y[11]首次提出一种基于卷

积神经网络的文档预测模型，提出文本深度学习方法。郭丽敏[12]构建出基于题名、关键词的深度神经网络模型，能够在大规模数据集上预测中图分类号。吴洁等[13]立足于专利分类，首先将图书转换成单词图，进而采用图卷积神经网络训练每个单词特征，最终计算整张图的表示。戎璐与张亚洲等[2]提出一种注意力序列到序列模型解决大规模层次图书分类任务，并分析该算法的时间性能，验证其实验性能。Zhang Yazhou 等[14]将预训练语言模型深度双向转换网络（Bidirectional Encoder Representations from Transformers，BERT）应用到大规模评论与文本对话分类任务中，在两个基准数据集上获得良好的实验结果。他们也提出一种先进的量子概率驱动式语言模型，应用到文档分类任务中，提出基于量子理论的文档分类方法[15]。

基于深度学习的方法是目前最前沿的图书自动分类方案，本研究也将延续这种模式。但是与已有方法截然不同的是，本研究将提示学习与大语言模型 ChatGPT 引入图书自动分类领域，构建图书自动分类新范式，即从现有的“大规模训练样本 – 训练模型 – 更新参数 – 分类”转变为“小样本 – 询问大语言模型 – 提示 – 生成类目”。具体如图 2 所示。



（a）基于预训练语言模型的提示学习路径



（b）基于ChatGPT模型的提示学习路径

图 2　基于多模态注意力提示学习的图书自动分类方案

2　研究方法

本部分将详述研究问题与多模态注意力提示学习网络方案。

2. 1　问题描述

本部分将描述研究对象、研究目标、研究流程与模型框架。

（1）研究对象：本研究将数据集中第 i 篇多模态图书作为研究目标 $D_i = \{T_i, V_i\}$，其中 T_i 表示文本信息，V_i 表示图像信息（即封面图）。此外，文本信息可以表示为 $T_i = \{H_i, K_i, J_i\}$，H_i 表示文本标题，K_i 表示关键词，J_i 表示简要信息。

（2）研究目标：按中图分类法，图书类目的一级、二级、三级类目 $L = \{l_1, l_2, l_3\}$。

（3）研究流程：（a）本研究基于预训练模型注意力提示学习网络（multi – modal attentive prompt learning，MAPL），将第 i 篇多模态图书的文本信息与图像信息分别表征为向量，再人工设计 5 种初始提示模板，记为 $p = \{P_1, P_2, P_3, P_4, P_5\}$，采用注意力机制生成最终多模态提示模板 $Mp = Attention(p)$，输入预训练语言模型 RoBERTa 中，进行一级、二级、三级类目分类，简写为 $\{l_1, l_2, l_3\} = MAPL(\{T_i, V_i\})$；（b）本研究提出一个基于 ChatGPT 大语言模型的提示学习模型，根据第 i 篇多模态图书的文本信息，人工设计 5 种提示模板，记为 $p = \{P_1, P_2, P_3, P_4, P_5\}$，将其分别作为 ChatGPT 的用户输入，ChatGPT 将生成 5 种一级、二级、三级类目答案，针对这 5 种答案执行最大投票机制，选出最终一级、二级、三级类目答案，简写为 $\{l_1, l_2, l_3\} = ChatGPT(\{T_i\})$。

（4）模型框架：本研究展示一个基于注意力提示学习网络的图书自动分类方案，包括两种具体实施路径：①基于预训练语言模型的注意力提示学习模型，即 $MAPL$；②基于 ChatGPT 大语言模型的提示学习模型，即 ChatGPT，如图 2 所示。MAPL 模型包含 4 个核心模块，分别是多模态特征提取模块、多模态提示模板生成模块、注意力机制模块与图书分类模块。具体而言，首先通过文本预训练语言模型 BERT 将第 i 篇图书的文本部分 $T_i = \{H_i, K_i, J_i\}$ 的标题、关键词与简介中的每个单词表征为词嵌入，共生成三个词嵌入序列 $E_d = \{w_1^d, w_2^d, \cdots, w_n^d\}$，其中 $d \in \{H, K, J\}$ 分别对应标题、关键词与简介内容，$n \in \{q, s, r\}$ 表示标题、关键词与简介中单词数量。整个文本信息可以表征为标题、关键词与简介词嵌入序列的拼接，即 $E_T = E_H \oplus E_K \oplus E_J$。其次，通过另一残差神经网络模型结合 Transformer 将第 i 篇图书的图像部分表征为视觉向量 $E_V = \{v_1^v, v_2^v, \cdots, v_n^v\}$。再次，人工设计出 5 个典型多模态提示模板，记为 $p = \{P_1, P_2, P_3, P_4, P_5\}$，利用注意力机制加权式融合以上模板，生成唯一多模态提示模板，$Mp = Attention(p)$。最后，将其输入预训练语言模型 RoBERTa 确定图书类目，达到自动图书分类的目的。ChatGPT 模型同样包含 4 个核心模块，分别是提示模板构造模块、ChatGPT API 调用模块、最大投票机制模块与图书分类模块。具体而言，首先根据第 i 篇多模态图书的文本信息，人工设计 5 种提示模板，记为 $p = \{P_1, P_2, P_3, P_4, P_5\}$。然后将其分别作为 ChatGPT API 的用户输入，调用 ChatGPT 接口将生成 5 种一级、二级、三级类目答案，针对这 5 种答案执行最大投票机制，根据投票结果，选出最终一级、二级、三级类目答案，达到自动分类的目的。从上述表述可以知道，两种实施方案均不涉及模型训练与参数更新迭代，只是需要用户（图书馆员）自由设计若干个提示模板即可完成自动分类。

2.2 基于预训练模型的注意力提示学习网络

注意力提示学习网络不同于传统深度神经网络，其通过向预训练模型提供特定的输入提示（prompt），让预训练模型生成符合特定条件的输出，而无需改变模型结构与参数，其被称为自然语言第四范式，已经开始应用于文本分类、文本生成等领域。

2.2.1 多模态特征提取

多模态特征提取旨在将图书的文本、图像编码表征为序列向量。

（1）对于图书的文本信息，每一篇图书的文本信息是由标题（H）、关键词（K）与简介（J）构成，本研究将通过预训练语言模型将此三者分别表征为向量。以标题为例，假设标题共有 q 个单词，本研究将其表征为 $H = \{w_1^H, w_2^H, \cdots, w_q^H\}$。鉴于预训练深度双向转换网络 $BERT$ 具有强大的特征表征能力，且在对中文进行建模时，不需要分词等优点，本研究采用深度双向转换网络 $BERT$ 将标题中每个单词嵌入为词向量（亦称词嵌入），形式转化为：

$$E_H = \{e_1^H, e_2^H, \cdots, e_q^H\} = BERT(H) \tag{1}$$

同样，本研究可以将文本部分的关键词、简介部分通过深度双向转换网络 BERT 表征为另外两个词向量序列，记为：

$$E_k = \{e_1^K, e_2^K, \cdots, e_s^K\} = BERT(K) \tag{2}$$

$$E_J = \{e_1^J, e_2^J, \cdots, e_r^J\} = BERT(J) \tag{3}$$

本研究将拼接标题、关键词、简介向量，生成统一向量，作为文本信息的表征，形式化为：

$$E_T = E_H \oplus E_K \oplus E_J \tag{4}$$

本研究已经将多模态图书的文本部分表征为向量 E_T。与现有方法类似，本研究认为标题、关键词与简介已经能够充分表达文本部分的语义。

（2）本研究通过残差神经网络模型 ResNet 将多模态图书的图像部分表征为向量。首先，将图像部分自上而下、自左向右划分为 n 个区域，表征为 $V = \{v_1^V, v_2^V, \cdots, v_n^V\}$。其次，本研究采用 *ResNet* 模型将每个区域表征为视觉向量。最后，为了保存区域之间的空间交互信息，将这 n 个区域向量输入预训练语言模型 Transformer 中，获得最终的图像向量：

$$E_V = \{e_1^V, e_2^V, \cdots, e_n^V\} = Transformer(ResNet(V)) \tag{5}$$

至此，本研究已经完成多模态图书的文本、图像两部分的特征表征。

2.2.2 多模态提示模板生成

在提示工程中，输入的语句被形式化为自然语言模板，而图书分类任务被转换为填空任务。该模板是当前任务的背景描述，标签词是由当前上下文中的预训练语言模型高概率预测的词语。然后将图书类目与标签词对应起来，从而完成图书分类。本研究将图书的多模态信息（文本与图像）用于构建提示模板，将一级、二级、三级类目视作标签词。

在提示学习中，生成多模态模板的过程可以分为以下 4 个步骤：

（1）定义多模态图书模板的结构，例如模板中应该包含哪些元素、元素之间的关系等。本研究将设计一个完形填空题目。

（2）确定模板中需要填充的内容，可能包括关键词、实体、标签等。本研究主要关注图书分类任务，模板内容包含文本、图像、提示语与[MASK]等元素，其中图像将作为辅助与补充信息与文本信息拼接，但是为了构造不同类型模板，图像与文本的位置可能会变化。

（3）设计填充内容的提示语，对于每个需要填充的内容，需要设计相应的提示语引导模型生成合适的标签词。本研究中，提示语是围绕图书分类任务引导模型判断图书类目的相关语句，例如“所以，该图书的一级类目是[MASK]”。

（4）根据上述步骤，生成最终的模板，即包含结构和填充内容提示语的完整文本。该文本将作为预训练语言模型的输入。

结合文本信息表征 E_T、图像信息表征 E_V，根据不同的提示语，本研究设计出 5 种典型多模态提示模板，如下所述（注：以一级类目分类为例）：

P_1：{文本：[E_T]. 图像：[E_V]. 依据《中图法》第五版，该图书的基本大类（一级类目）是[MASK]. }

P_2：{图像：[E_V]. 文本：[E_T]. 依据《中图法》第五版，该图书的基本大类（一级类目）是[MASK]. }

P_3：{文本：[E_T]. 图像：[E_V]. 依据《中图法》第五版，请问该图书隶属于哪个基本大类（一级类目)? [MASK]. }

P_4：{图像：[E_V]. 文本：[E_T]. 依据《中图法》第五版，请问该图书隶属于哪个基本大类（一级类目)? [MASK]. }

P_5：{多模态图书：[E_T] [E_V]. 依据《中图法》第五版，请问该图书隶属于哪个基本大类（一级类目)? [MASK]. }

面对图书类目二级、三级标签分类时，本研究将一级类目的预测信息作为辅助信息，输出二级或三级类目，如下所示（注：以二级类目分类为例）：

P_1：文本：[E_T]. 图像：[E_V]. 依据《中图法》第五版，该图书的基本大类（一级类目）是[MASK]，在此基础上，该图书的二级类目是[MASK].

P_2：图像：[E_V]. 文本：[E_T]. 依据《中图法》第五版，该图书的基本大类（一级类目）是[MASK]，在此基础上，该图书的二级类目属于[MASK].

P_3：文本：[E_T]. 图像：[E_V]. 依据《中图法》第五版，该图书的基本大类（一级类目）是[MASK]，请问该图书的二级类目是?[MASK].

P_4：图像：[E_V]. 文本：[E_T]. 依据《中图法》第五版，该图书的基本大类（一级类目）是[MASK]，请问

该图书的二级类目是?［MASK］.

P_5：多模态图书：［E_T］［E_V］. 依据《中图法》第五版，该图书的基本大类（一级类目）是［MASK］，请问该图书的二级类目是?［MASK］.

值得说明的是，真实的提示输入应当是这些模板的向量序列，即文本嵌入、图像嵌入、提示语嵌入、［MASK］嵌入，以及两个开始、结尾位置嵌入［CLS］、［SEP］。真实的提示输入可以形式化为：［E（［CLS］），E（｛P_1，P_2，P_3，P_4，P_5｝），E（［$MASK$］），E（［SEP］）］。

2.2.3 多提示注意力机制

本研究基于专业经验与领域知识设计出 5 种多模态提示模板 p = ｛P_1，P_2，P_3，P_4，P_5｝，尽管如此，仍然可能存在不合理或没有贡献的提示模板。为了解决潜在的缺陷，本研究提出多提示注意力机制去控制每种模板的贡献，通过衡量每种模板的权重生成一个加权式融合提示模板 Mp。

具体而言，本研究将 P_1 视作查询模板，即 $Q_1 = W_q P_1$，将其他 5 个（包括自己）提示模板视作键与值，即 $K_j = W_k P_j$，$U_j = W_u P_j$，其中 W_q、W_k 与 W_u 都是权值参数，且 $P_j \in$ ｛P_1，P_2，P_3，P_4，P_5｝。那么注意力分数可以计算为：

$$\alpha_j = softmax\left(\frac{Q_1 K_j}{\sqrt{d_k}}\right)U_j = softmax\left(\frac{W_q P_1 W_k P_j}{\sqrt{d_k}}\right)W_u P_j \tag{6}$$

$$Mp = \sum_{j=1}^{5} \alpha_j P_j \tag{7}$$

2.2.4 图书分类

本研究已经获得加权式多模态提示模板 Mp，将其输入预训练语言模型 RoBERTa 中，通过“询问” RoBERTa 获得提示模板中［MASK］位置的类目预测值。本研究选择 RoBERTa 作为预训练语言模型的原因是其使用更多的文本数据进行训练，训练时间更长，能够更好地学习语言的规律和特点。鉴于 RoBERTa 已经由原作者训练完毕，图书分类过程无需再次训练该模型，不涉及参数更新，RoBERTa 将输出［MASK］位置的类目的预测概率，拥有最高预测概率的是类目。整个分类过程可以形式化为：

$$Label = max\left(\frac{e^{ROBERTa(l \mid Mp)}}{\sum_i e^{ROBERTa(i \mid Mp)}}\right) \tag{8}$$

其中，$Label$ 表示类目，$e^{ROBERTa(l \mid Mp)}$ 表示预测第 l 个类目的概率。

2.3 基于大语言模型 ChatGPT 的提示学习网络

本研究通过人工设计提示模板，调用 ChatGPT API 接口，生成符合特定条件的输出。

2.3.1 提示模板构造

为激发大语言模型的潜能，本研究根据图书的文本信息（ChatGPT 只能接受文本信息），依据问答与填空两种提示模式，构造 5 种不同类型的提示模板，作为 ChatGPT 接口的输入信息，如下所述（注：以一级类目图书分类为例）：

P_1：已知图书的标题（H）、关键词（K）、简介（J）。依据《中图法》第五版，该图书的一级类目是［MASK］。

P_2：已知图书的标题（H）、关键词（K）、简介（J）。结合《中图法》第五版、郑州图书馆、广州图书馆分类准则，该图书的一级类目是［MASK］。

P_3：已知图书的标题（H）、关键词（K）、简介（J）。依据《中图法》第五版，请问该图书隶属于哪个基本大类（一级类目）?［MASK］。

P_4：已知图书的标题（H）、关键词（K）、简介（J）。结合《中图法》第五版、郑州图书馆、广州图书馆分类准则，请问该图书隶属于哪个一级类目?［MASK］。

P_5：根据图书的标题（H）、关键词（K）、简介（J）。将其归结为哪种一级类目最合适?［MASK］.

面对图书类目二级、三级标签分类时，本研究将一级类目的预测信息作为辅助信息，输出二级或三级类目，如下所示（注：以二级类目分类为例）：

P_1：已知图书的标题（H）、关键词（K）、简介（J），以及基本大类（一级类目）信息。依据《中图法》第五版，该图书的二级类目是［MASK］。

P_2：已知图书的标题（H）、关键词（K）、简介（J），以及基本大类（一级类目）信息。结合《中图法》第五版、郑州图书馆、广州图书馆分类准则，该图书的二级类目属于［MASK］。

P_3：已知图书的标题（H）、关键词（K）、简介（J），以及基本大类（一级类目）信息。依据《中图法》第五版，请问该图书的二级类目是?［MASK］。

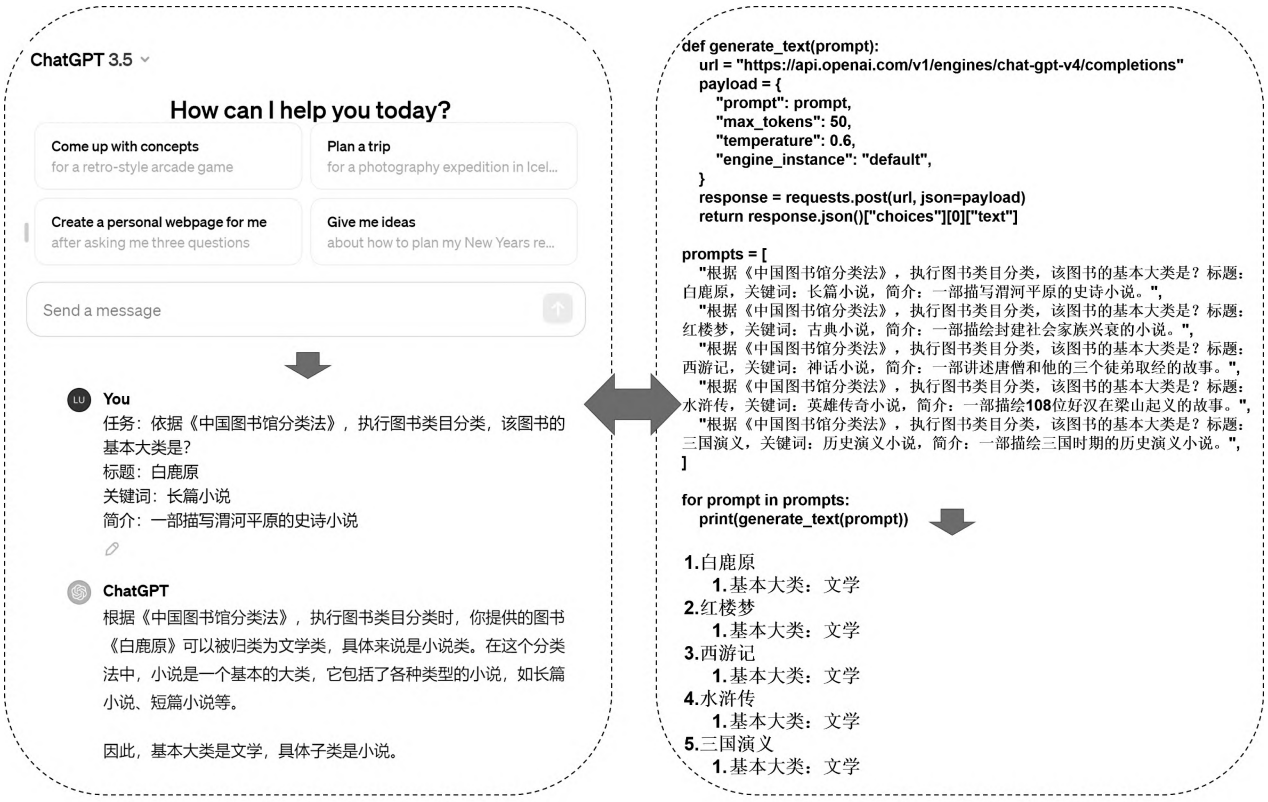
P_4：已知图书的标题（H）、关键词（K）、简介（J），以及基本大类（一级类目）信息。结合《中图法》第五版、郑州图书馆、广州图书馆分类准则，请问该图书的二级类目是?［MASK］。

P_5：根据图书的标题（H）、关键词（K）、简介（J），以及基本大类（一级类目）信息。将其归结为哪种二级类目最合适?［MASK］。

值得说明的是，代码实现时，调用 generate_ text 函数里传入的用户提示是一个列表，列表里包含 5 个一级类目提示或者二级、三级类目提示。

2.3.2 ChatGPT API 调用

OpenAI 官方网站提供两种使用 ChatGPT 的方式：第一种是注册账号后进入 ChatGPT 体验网站后，直接在输入框内输入用户提示模板，ChatGPT 将根据用户提示输入迅速生成答案并显示到界面上；第二种是根据 OpenAI 公布的接口，跳转到 API key 界面，点击 Create new secret key 生成 API Keys，然后利用 Pip 工具安装 OpenAI 应用程序编程接口的 Python 包。设计程序调用 completion. choices［0］. message 即可显示用户自己的提示输入与 ChatGPT 生成的答案。为了能够一次性输入 5 个用户提示，该程序可以将这 5 个用户提示封装到一个 prompts 列表，然后循环迭代每个提示，并调用 generate_ text 函数来生成文本，这样即可实现一次性传入 5 种用户提示，获得对应的类目答案，如图 3 所示。通过上述方式即可用该程序完成 ChatGPT API 的调用，并获得 5 种用户提示对应的图书类目答案。



（a）直接使用网站完成图书分类　　（b）调用API接口完成图书分类

图 3 直接使用网站与调用应用程序接口对比

2.3.3 最大投票机制

最大投票机制（也称为最大多数投票机制）是一种简单的集成学习技术，其中决策结果是通过获得最多选票的候选选项来确定的，允许多个基本学习器的预测结合。在这种机制中，假定每个基础学习器都是投票者，每个类目都是竞争者，然后类目获得最多选票的那个将被选为获胜者。

本研究通过调用 ChatGPT API 可以获得 5 个关于图书类目的答案，假设答案集中于 A、B 两种类目，直接计

算 A、B 类目对应的数量，数量最多者即被认定为最终的图书类目。根据同样的方式，可以选出最终一级、二级、三级类目答案，达到自动分类的目的。总结而言，第二种图书自动分类路径同样无需模型训练与参数更新，不需要建立大规模训练样本集，可以直接利用 ChatGPT 的提示学习能力，无需了解其内部机制。

3 实证检验

本部分将描述实验步骤、实验设置，通过与一系列前沿方案对比，验证本方案的有效性。

3.1 数据选取

本研究实证检验所用的数据来源于两所国家一级图书馆的馆藏，即广东省广州图书馆和河南省郑州图书馆部分藏书资源。通过在广州/郑州图书馆网站页面“高级检索”－“分类导航”－“中图分类”按照 10 个一级类目检索浏览，获取每个所检索的图书的标题、主题（关键词）、内容简介等文本信息以及伴随的封面图。对于缺少封面图或关键词等信息的图书，将直接弃用。为了验证本研究方案对小样本图书自动分类的有效性，本研究从两所图书馆数据中各自选择 100 本图书（其中每类图书正好 10 种），共计 200 种。

本研究额外建起一个大规模图像分类样本集，此时不再强制要求图书具有封面图，原因是广州图书馆许多书籍并未上传封面。本研究根据图书馆中的“中图分类浏览”标准按类目收集数据，最终得到 114 823 本图书。值得强调的是，本研究收集小样本数据集的原因并不是因为提示学习只能应用于小规模图书分类场景，而是为了展示提示学习与传统深度学习不同，无需建立大规模训练数据集，即可学习到图书规律。提示学习的初衷正是为了解决大规模任务而生。本研究将在第二个大规模数据集上验证提示学习的有效性与先进性。表 1 展示小样本与大规模数据集样本统计特性。

表 1 小样本与大规模数据集样本统计

一级类目	小样本数据集				大规模数据集			
	训练集	验证集	测试集	样本总数	训练集	验证集	测试集	样本总数
政治法律	12	4	4	20	4 352	1 450	1 450	7 252
音乐	12	4	4	20	10 015	3 338	3 338	16 693
历史	12	4	4	20	6 952	2 197	2 197	10 986
地理	12	4	4	20	11 234	3 744	3 744	18 722
军事	12	4	4	20	7 535	2 511	2 511	12 557
经济	12	4	4	20	7 730	2 576	2 576	12 883
航空航天	12	4	4	20	3 237	1 078	1 078	5 393
文学	12	4	4	20	10 410	3 469	3 469	17 348
艺术	12	4	4	20	6 712	2 237	2 237	11 186
数学	12	4	4	20	1 083	360	360	1 803

3.2 实验设置

（1）评价指标。本研究将计算精准率（precision，P）、召回率（recall，R）、F1 数值作为评准依据。同时，也将准确率（accuracy，Acc）与 F1 用于模型剥离实验等后续分析。每种评价指标定义如下：

$$Precision = \frac{N_{tp}}{N_{tp} + N_{fp}} \tag{9}$$

$$Recall = \frac{N_{tp}}{N_{tp} + N_{fn}} \tag{10}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{11}$$

$$Accuracy = \frac{N_{tp} + N_{tn}}{N} \tag{12}$$

其中，N_{tp}、N_{fp}、N_{fn}分别表示正类预测为正类、负类预测为正类、正类预测为负类的数量。N 是总样本数量。

（2）超参数设置。本研究采用 PyTorch 工具搭建多模态提示学习网络模型，硬件环境是 GeForceTeslaA100

40G GPU。文本向量选择由 BERT 表征初始化，图像向量由 ResNet50 初始化。

3.3 对比方案

本研究使用前沿机器学习、深度学习方法实现图书自动分类，以便与多模态提示学习网络模型对比。其中，支持向量机 SVM 与朴素贝叶斯 NB 机器学习算法被选为两种比较目标。分类流程是：利用词嵌入技术提取标题、关键词与简介特征，拼接后输入机器学习分类器。

深度学习方面，RNN、CNN、LSTM 以及基于注意力机制的 LSTM 模型（简称 Att－LSTM）、预训练语言模型 BERT 等深度神经网络将参与比较。具体而言，利用包含 1.1 亿个参数的“BERT－Base，Chinese”预训练模型技术提取标题、关键词与简介特征，将三者拼接成图书特征，输入深度神经网络，经过训练收敛后，通过 softmax 函数获得图书类目。

多模态方面，将文本预训练语言模型 BERT 与残差神经网络模型 ResNet 分别表征文本、图像特征，拼接成多模态特征后，利用 softmax 函数获得图书类目。

3.4 小样本数据集实证结果

小样本图书分类方案的实证结果如表 2 所示。对于一级类目，可以观察到机器学习图书分类方法反而胜过深度学习方法，例如 CNN、LSTM、RNN 等，原因是深度学习算法严重依赖大规模数据集，在小样本数据集上出现过拟合，因此性能严重下降。相对而言，机器学习算法可以适应数百条规模数据集，因此可以正常训练与分类。但是，本研究同样测试机器学习算法在大规模（11 万条）数据集上的性能，发现其性能高于本研究的结果（见 3.5 部分），这也证实机器学习比较依赖数据集。此外，可以观察朴素贝叶斯算法 F1 结果仅为 60.0%，支持向量机 SVM 大幅度优于朴素贝叶斯算法。卷积神经网络与循环神经网络取得同等级实验结果，而长短期记忆网络 LSTM 因为能建模长期词序信息，在各项指标上得到较好的提升。基于注意力机制的 LSTM 网络因其能够权衡各个单词之间的重要程度，超越了标准长短期记忆网络。深度双向转换网络 BERT 凭借其预训练模型强大的特征提取能力，F1 分值大幅度超越 Att－LSTM、LSTM、CNN 等主流神经网络，达到 64.2%。BERT 表现优良的原因是其模型包括深层双向编码以及自注意力机制捕捉句内远距离依赖。

表 2 小样本数据集各种方案实证结果

数据集	模型	一级类目			二级类目			三级类目		
		P	R	F1	P	R	F1	P	R	F1
小样本数据集	SVM	0.631	0.635	0.632	0.504	0.516	0.511	0.377	0.381	0.380
	NB	0.598	0.604	0.600	0.497	0.495	0.497	0.360	0.357	0.359
	CNN	0.421	0.429	0.425	0.410	0.415	0.412	0.350	0.369	0.366
	RNN	0.512	0.522	0.516	0.521	0.526	0.524	0.401	0.407	0.405
	LSTM	0.517	0.523	0.520	0.520	0.534	0.531	0.420	0.425	0.422
	Att－LSTM	0.529	0.533	0.531	0.523	0.540	0.534	0.431	0.432	0.431
	Att－LSTM（大规模数据集）	0.741	0.754	0.741	0.643	0.654	0.650	0.482	0.488	0.484
	BERT	0.640	0.645	0.642	0.629	0.622	0.624	0.450	0.455	0.453
	BERT（大规模数据集）	0.775	0.780	0.777	0.701	0.698	0.698	0.492	0.493	0.491
	多模态方案	0.721	0.727	0.724	0.677	0.674	0.675	0.477	0.482	0.479
	多模态注意力提示学习网络 MAPL	0.795	0.810	0.802	0.733	0.729	0.731	0.501	0.504	0.502
	ChatGPT	0.915	0.909	0.907	0.807	0.791	0.802	0.702	0.701	0.701

为了解决深度学习算法过拟合问题，本研究基于注意力机制的 LSTM 网络模型 Att－LSTM 与 BERT 模型在大

规模数据集（表2，11万条数据）上重新训练运行，两者的性能取得巨大提升，性能提升比达到50%以上。这项实验更加证实：深度学习图书分类方法取得性能优势的前提是拥有大规模图书数据集。本研究提出的基于多模态注意力提示学习模型MAPL超过BERT（大规模），在各项指标上获得最佳分类性能，F1分值达到80.2%，相比于BERT方法提升5.2%。实证结果表明本研究提出模型的有效性。进一步地，本研究第二套实施路径ChatGPT获得卓越性能表现，刷新了图书自动分类的准确率，达到90.7%，显著性超越MAPL达11.3%，相比BERT增幅达到16.2%，相比Att－LSTM增幅达到24.2%，究其原因是其以拥有上万亿参数的GPT－3.5作为基座模型，大语言模型在推理能力、理解能力方面实现质的飞跃。

对于二级、三级图书分类任务，可以观察到全部方案性能均呈现明显的下降趋势，原因在于二级与三级类目数量的指数级增加，从而导致细粒度图书分类的挑战性提升。而BERT（大规模）凭借其预训练语言模型与1.1亿参数的优势，胜过其他深度学习方法。本研究提出的MAPL方案表现最佳，F1结果分别是73.1%与50.2%，增长幅度分别是6.25%与4.16%，原因在于MAPL不仅利用了预训练语言模型的已有知识，在生成二级、三级标签时能够利用一级、二级标签信息。相比于其他模型，ChatGPT在二级、三级类目上表现也极为突出，在二级类目上突破了80%，在三级类目上突破70%，远远超越其他模型。更重要的是，ChatGPT采用基于人类反馈的强化学习方式进行训练，使得模型生成的文本更符合人类的期待。这样可以实现与图书馆员的交互，图书馆员可以随时对ChatGPT反馈与引导，例如对其每次分类结果进行打分，使其能够从指令集合中采样一些新的指令作为输入数据，并结合打分信息，使用强化学习算法升级优化自身性能。此外，ChatGPT与MAPL无需模型训练即可完成图书分类任务，实时性非常强，特别是ChatGPT可以在50秒完成对测试集40本书的分类，且正确率达到90%左右，有潜力成为图书馆员的辅助工具。

本研究提出的两种提示学习方案完全可以应用于小样本数据集，为图书馆员带来两点明显好处：（1）无需图书馆员收集大规模训练数据集，减轻其工作压力；（2）无需将模型再次训练，可以直接应用，降低方案使用难度。

3.5 大规模数据集实证结果

大规模图书分类方案的实证结果如表3所示。可以观察到所有模型在大规模数据集上性能均取得大幅度提升，原因是机器学习与深度学习依赖于大规模训练样本。对于一级类目，可以观察朴素贝叶斯算法实证结果最差，F1结果仅为66.0%。支持向量机SVM大幅度优于另外一种算法，但是又低于卷积神经网络。这表明深度学习算法在大规模数据集上相比于机器学习能取得更优的结果。卷积神经网络与循环神经网络取得同等级实验结果，而长短期记忆网络LSTM因为能建模长期词序信息，在各项指标上取得较好提升。基于注意力机制的LSTM网络因其能够权衡各个单词之间的重要程度，超越了标准长短期记忆网络。深度双向转换网络BERT凭借其预训练模型强大的特征提取能力，F1分值超越Att－LSTM、LSTM、CNN等主流神经网络，达到77.7%。本研究提出的MAPL超过BERT，在各项指标上获得最佳分类性能，F1分值达到81.5%，相比于BERT方法提升4.2%。MAPL采用的RoBERTa预训练语言模型与BERT具有等量的参数量，因此提升幅度没有ChatGPT那么大。而ChatGPT的显著性仍然超越其他方案，在一级、二级、三级类目任务中获得最佳性能，分别达到91.1%、80.8%、73.0%。ChatGPT在大规模图书分类场景与其在小样本数据集上性能接近，说明其可靠性，既可以适用于小样本学习，也可以应用于大规模图书分类场景。

表3 大规模数据集各种方案实证结果

数据集	模型	一级类目			二级类目			三级类目		
		P	R	F1	P	R	F1	P	R	F1
大规模数据集	SVM	0.711	0.715	0.713	0.614	0.626	0.621	0.464	0.471	0.467
	NB	0.658	0.664	0.660	0.567	0.555	0.560	0.430	0.437	0.434
	CNN	0.714	0.721	0.719	0.620	0.621	0.622	0.459	0.471	0.467
	RNN	0.726	0.736	0.726	0.631	0.639	0.637	0.461	0.465	0.464
	LSTM	0.737	0.743	0.739	0.630	0.644	0.641	0.470	0.476	0.473
	Att－LSTM	0.741	0.754	0.741	0.643	0.654	0.650	0.482	0.488	0.484

数据集	模型	一级类目			二级类目			三级类目		
		P	R	F1	P	R	F1	P	R	F1
大规模数据集	BERT	0.775	0.780	0.777	0.701	0.698	0.698	0.492	0.493	0.491
	多模态方案	0.752	0.758	0.754	0.691	0.697	0.693	0.487	0.492	0.489
	多模态注意力提示学习网络 MAPL	0.811	0.817	0.815	0.752	0.744	0.747	0.535	0.524	0.529
	ChatGPT	0.911	0.914	0.911	0.812	0.804	0.808	0.734	0.727	0.730

3.6 图书消融分析

为了研究目前方案中多模态、提示学习与注意力机制对分类性能的贡献度，本研究继续开展消融实验，遵循一次去除一个组件的规则，分别设计了3个子方案：（1）无多模态，去除图像信息，只用文本信息；（2）无提示学习，去除提示模板等内容；（3）无注意力，去除注意力机制，从提示模板中，随机挑选一个输入预训练语言模型RoBERTa，实证结果如表4所示。

表4 方案消融实证结果

数据集	模型	一级类目		二级类目		三级类目	
		Acc	F1	Acc	F1	Acc	F1
小样本数据集	无多模态	0.777	0.780	0.695	0.693	0.490	0.490
	无提示学习	0.731	0.730	0.681	0.682	0.487	0.483
	无注意力	0.794	0.792	0.719	0.717	0.497	0.496
	MAPL	0.805	0.802	0.734	0.731	0.505	0.502
	随机提示	0.898	0.904	0.784	0.788	0.674	0.677
	ChatGPT	0.907	0.910	0.802	0.801	0.701	0.713
大规模数据集	无多模态	0.774	0.781	0.717	0.723	0.501	0.508
	无提示学习	0.752	0.757	0.703	0.712	0.492	0.493
	无注意力	0.801	0.808	0.731	0.736	0.515	0.522
	MAPL	0.815	0.821	0.747	0.750	0.529	0.535
	随机提示	0.900	0.901	0.788	0.796	0.714	0.721
	ChatGPT	0.911	0.919	0.808	0.813	0.730	0.742

可以观察到，首先，在一级类目分类中，无提示学习取得最差实证结果，根据消融原则，表明去除提示学习对分类方案负影响最大，即提示学习对本研究提出的MAPL方案贡献度最高。其次，无注意力取得最佳分类结果，表明注意力机制对MAPL方案贡献度最低，占据次要地位。无多模态方案的性能低于无注意力，高于无提示学习，说明多模态信息占据中等地位。这也印证了人工智能领域的共识：多种模态将提供更丰富、更准确的信息描述。二级类目分类结果，各个方案分类趋势等同于一级类目分类，提示学习仍然对分类性能影响最高。在三级类目分类任务中，三种方案取得类似结果，表明三者均无法有效处理高难度子类目分类。上述三个子方法均比MAPL实证结果差，说明多模态、提示学习、注意力机制对提升分类模型都有必要。

鉴于ChatGPT模型原理与代码并未开源，因此，本研究的消融分析主要关注提示模板消融分析。

3.7 一级类目分类结果

本文主要调查MAPL与ChatGPT在大规模数据集上的实证结果（表5）。可以观察到，MAPL方案在数学、地理、音乐等类目分类结果最佳，F1分值全部超越90%。其次，文学、历史也有良好的分类结果，F1分值在80%左右。相比而言，MAPL方案在政法、经济、艺术、军事与航空一级类目上表现不佳，原因主要是这些类目的图书在标题或关键词领域均有与其他类目相关的词语，导致方案出现预测失误。ChatGPT表现较为均衡，未出

现大幅度下降情况，但是在政法、经济、文学、历史与军事类目低于平均值，在音乐、地理、数学、航空方面表现良好。

表 5 一级类目实证结果

MAPL	类目	政法	音乐	经济	文学	艺术	历史	地理	数学	军事	航空
	Acc	0. 746	0. 901	0. 788	0. 805	0. 776	0. 801	0. 913	0. 942	0. 764	0. 574
	F1	0. 741	0. 904	0. 792	0. 804	0. 773	0. 795	0. 917	0. 940	0. 769	0. 572
ChatGPT	类目	政法	音乐	经济	文学	艺术	历史	地理	数学	军事	航空
	Acc	0. 892	0. 935	0. 882	0. 909	0. 920	0. 906	0. 961	0. 945	0. 904	0. 922
	F1	0. 890	0. 927	0. 876	0. 908	0. 914	0. 905	0. 954	0. 945	0. 903	0. 917

3. 8 各方案分类时间对比

本研究提出的方案具有时效性强的优势。为了验证该观点，开展各方案分类时间对比，即统计每一种对比方案的分类时间，包括数据预处理、模型训练、分类等过程。具体而言，本研究将调用 Python 语言内置的时间函数，统计每个方案的开始时间与结束时间，从而计算出每个方案分类时间，对比结果如图 4 所示。

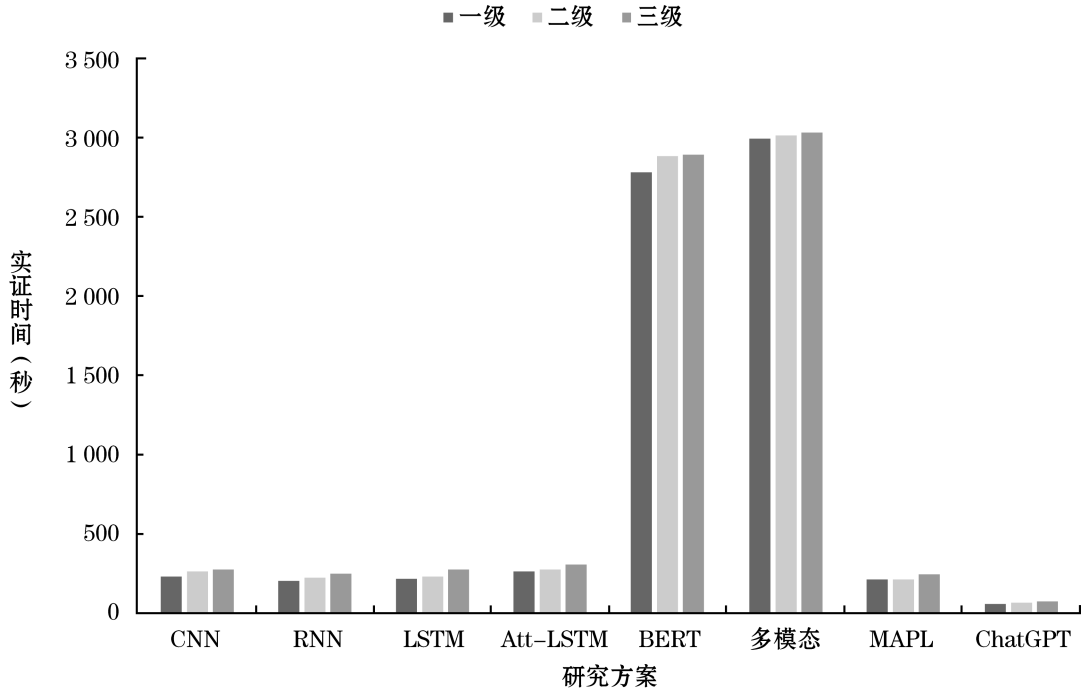


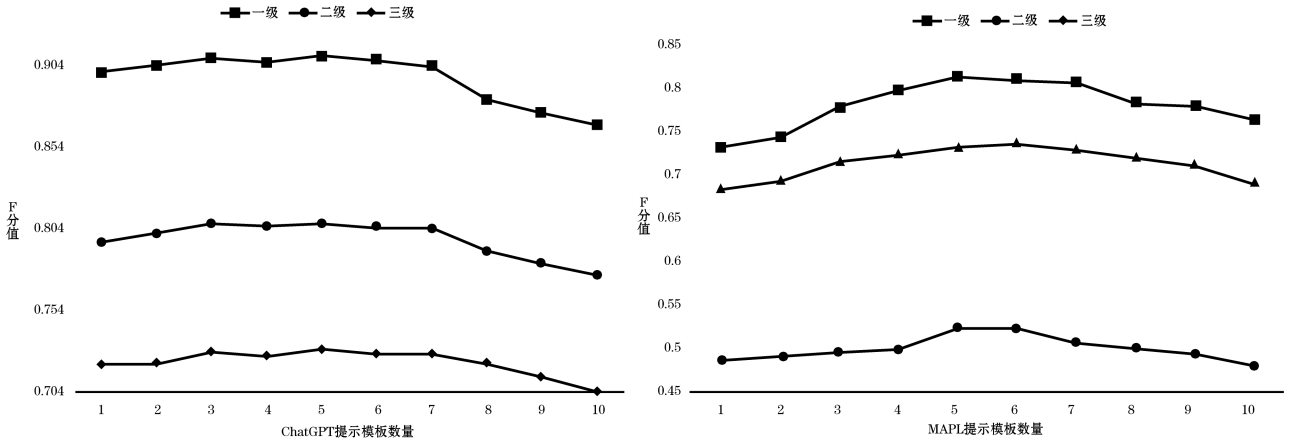
图 4 实证时间结果

我们可以观察到，多模态方案分类时间最长，原因是其需要分别训练文本预训练语言模型 BERT 与残差神经网络模型 ResNet，两者都拥有庞大的参数量。BERT 含有 1. 1 亿参数，ResNet 拥有 118 万参数。RNN 取得最短分类时间，原因是其不需要卷积运算，也不涉及预训练语言模型，因此在小样本数据集上可以很快完成训练。具体而言，在一级类目分类上，本研究提出的 MAPL 方案取得第三短分类时间，其时间主要耗费于其调用大规模预训练语言模型。在训练阶段，MAPL 可以明显缩短时间，最终在分类阶段超越 CNN、LSTM、Att－LSTM 等其他深度学习方案。虽然本方案提出的 MAPL 比 RNN 耗时多 15 秒，但是分类性能显著性超越 RNN 方案，增长幅度达 60％。此外，在二级、三级类目分类上，由于任务愈趋复杂，各方案分类时间均增长。但是，本研究 MAPL 方案可以利用一级、二级的分类结果作为辅助信息，最终在全部方案中取得第二短时间，该结果证明 MAPL 方案可以在短时间内取得高分类效率。与之对比，ChatGPT 的时间优势更加凸显，可以在 50 秒左右完成对 40 本图书的分类（通过调用 API 方式），取得 90％的正确率。在二级、三级更加复杂的分类任务上，耗时接近，这得益于 GPT

-3.5 强大的基础能力。结果说明：（1）提示学习可以避免模型训练时间，因而运行时间更短；（2）ChatGPT 无论在时间还是性能方面都已经走在图书分类方法的前沿。

3.9 提示模板数量研究

为了研究提示模板数量对分类方案的影响，以及探索获得最佳分类性能所需模板数量，本研究开展提示模板数量实证研究，通过设计不同数量的提示模板，验证其在图书一级类目分类任务上的表现。具体如图 5 所示。



（a）提示模板数量对ChatGPT影响　　（b）提示模板数量对MAPL影响

图 5　提示模板数量对两种方案分类性能影响

在 ChatGPT 中，当提示模板数量从 1 逐步增加到 10，结果如图 5（a）所示。我们可以观察到，当提示模板数量没有超过 7 时，ChatGPT 对提示模板数量并不敏感，不同数量的提示模板对性能影响不大。但是当提示模板数量超过 7 后，ChatGPT 方案的性能开始出现明显下降，原因是当 ChatGPT 方案过多时，误差会相应增多，导致 ChatGPT 出现误判情况。

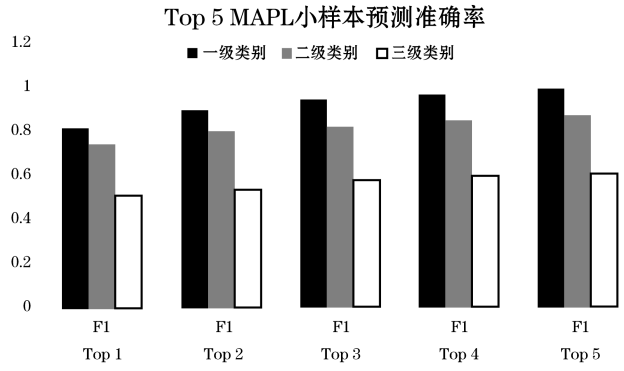
在 MAPL 中，当提示模板数量从 1 逐步增加到 10，结果如图 5（b）所示。我们可以观察到，当提示模板数量达到 6 时，MAPL 分类性能最优；当提示模板数量为 5 或者 7 时，MAPL 分类性能达到次高值。随着提示模板数量的增加，MAPL 性能呈现一种先上升后下降的总体趋势，这说明提示模板过多或者过少都会限制 MAPL 发挥最大优势。当提示模板过少时，无法有效地将分类任务转换为完形填空；当提示模板过多时，将引入更多的噪声；当提示模板数量达到 5、6、7 时，MAPL 在引入知识与噪声之间达到平衡。

3.10 TOP *N* 概率分析

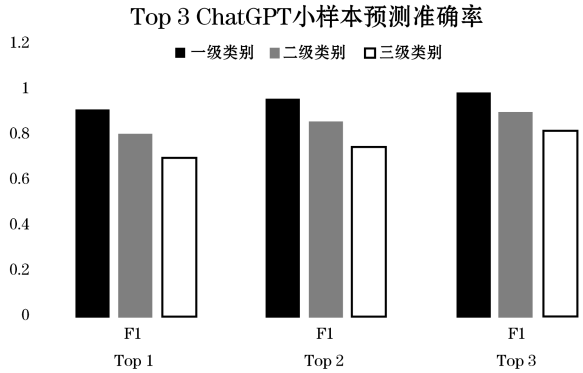
将 MAPL 与 ChatGPT 两个研究方案按照各个类目的预测概率，采纳最高者作为输出标签。本研究之前的数据仅仅显示了预测概率最高的类目的 F1 分值（图 6）。本研究展示出两者分别在小样本与大规模数据集上一级、二级、三级类目预测任务及排名前 5 个（或前 3 个）类目集合中包括正确类目的概率。例如，“历史”这一类目并未对应最高概率，可能对应次高概率。ChatGPT 生成的答案更加集中，因此只考虑前 3 个预测类目。

针对小样本数据集，在一级类目任务上，ChatGPT 可以在 Top3 达到 98.6%的预测准确率，而 MAPL 在 Top3 的 F1 值已经达到 93.4%，Top4 与 Top5 持续提升，最终达到 98.1%结果，这表明 MAPL 在前 5 个预测标签里包含真实标签的准确率达到 98%，而 ChatGPT 在前 3 个预测类目即可达到 98.6%。在二级、三级类目任务上，两者的 F1 结果也在不断提升。MAPL 的 Top5 的 F1 值为 85.9%、60.7%，相比于 Top1 提升了 17.5%、20.9%。而 ChatGPT 的 Top3 的值达到 90.6%与 81.4%，相比于 Top1 提升了 12.4%、16.3%。

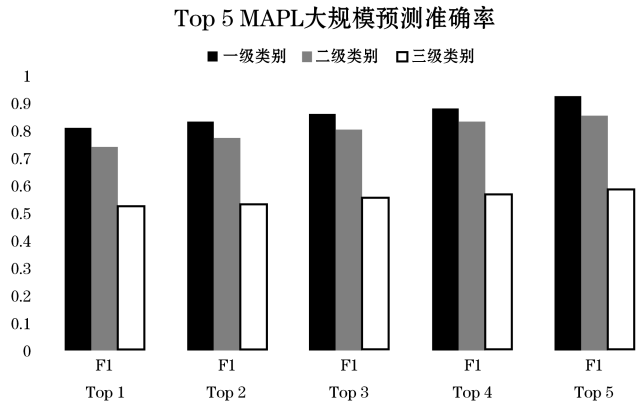
针对大规模样本集，在一级类目任务上，MAPL 在 Top5 达到 93.6%的预测准确率，而 ChatGPT 在 Top3 的 F1 值达到 96.6%，这表明 ChatGPT 对数据量相对不敏感，可靠性较高，在大规模数据集上仍能够在前 3 个标签获得非常高的准确率。在二级、三级类目任务上，两者的 F1 结果也在不断提升。最终，ChatGPT 在大规模样本集上，F1 值达到 88.6%与 78.4%，而 MAPL 的 Top5 预测准确率为 85.3%与 59.4%，稍逊于两者在小样本数据集上的表现。



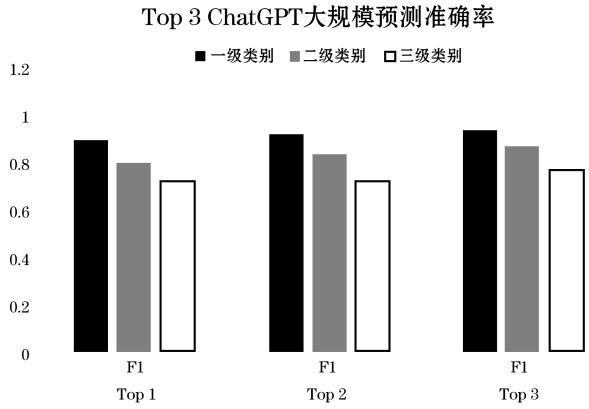
（a）MAPL小样本数据集Top5预测准确率



（b）ChatGPT小样本数据集Top3预测准确率



（c）MAPL大规模数据集Top5预测准确率



（d）ChatGPT大规模数据集Top3预测准确率

图 6　MAPL 与 ChatGPT 模型在小样本与大规模数据集上预测准确率

3. 11　局限性分析

以上结果已经验证本研究方案的有效性、简洁性、时效性。为了探索提示学习在图书馆学的潜力，分析目前方案存在的局限性，本研究尝试提供解决方案。详细讨论如下：

针对 ChatGPT 实施方案：（1）当前的 ChatGPT 仅局限于文本信息，无法处理多模态信息。但 OpenAI 刚刚发布的 GPT－4，已经可以处理多模态信息。本研究将会在未来工作中继续探索验证 GPT－4 的能力。（2）当前 ChatGPT 对国内封闭，虽然已经有大量研究者在使用测试它，但是仍无法满足国内需求。针对此，百度公司刚刚发布的“文心一言”大语言模型可以缓解该问题。（3）ChatGPT 方案无法与现实世界相接触。作为大型语言模型，它无法实时与外部世界互动，也无法利用计算器、数据库、搜索引擎等外部工具，导致它的知识也相对落后。本研究将在未来工作中引入搜索引擎工具。

针对 MAPL 实施方案：（1）对于一级、二级类目分类任务，尚不足以完全替代人工标注，无法全面满足图书馆智能化建设需求。但是，从图 6 可以观察到，一级、二级 Top5 的准确率已经分别达到 98. 1％与 85. 9％。在实际使用中，可以让 MAPL 输出前 5 个类目，借此可以帮助图书馆员快速定位，完成标注。（2）目前方案仍采用硬提示的方式，即凭借专业知识设计出提示模板，而不同的提示模板对分类性能影响较大。因此，后续改进方案将考虑采用软提示方式，即避免人工设计提示模板，而是让模型自己学习出一套模板，从而保证方案的“黑盒子”运行。（3）研究方案的单一性，仅能够满足图书分类任务，不具备多种任务统一性。图书馆智能化建设的重要体现在于全方位协同管理职能。优良的方案应当既完成图书分类，也可以具备其他能力，例如为读者推荐感兴趣的读物、实现热门书籍的排行预测、图书检索等。针对此，未来研究方案可以将多任务学习引入提示学习，形成多任务提示学习新范式。

3. 12　案例研究

为了挖掘提示学习的潜力，本研究分析 MAPL 方案错误分类案例，以便归纳模型缺点，进一步改善模型。表 6 展示出若干典型误分类案例。值得说明的是，尽管 ChatGPT 技术的封闭性使得研究者们无法探索其未来潜

力，但是本研究仍然展示若干误分类样例，旨在帮助图书馆领域理解 ChatGPT 的优缺点（表 7）。

表 6　MAPL 错误分类案例

数目	图　　书	类目	
		真实	预测
1	标题：璀璨文化：中国音乐的发展历史与现代传承 关键词：清商乐 摘要：本书主要围绕中国音乐的发展历史与现代传承展开分析。	文学	音乐
2	标题：近代民本思想转型史 关键词：哲学，民本思想 摘要：本书以中国传统民本思想的发展脉络和内在价值为主线，论述民本思想在近代转型期间，面对西方平等、自由、权利等观念的冲击而作出的坚守、改变与选择的思想过程。	政治	历史
3	标题：中国战略性新兴产业发展报告 关键词：新兴产业，产业发展 摘要：本书以新发展理念、新发展格局为战略指引，为形成一批具有全球竞争力的世界级产业集群提供咨询建议。	经济	军事
4	标题：航空技术与空中作战 关键词：航空，技术，空战 摘要：本书将人类百余年航空技术发展和空中作战历史，划分为活塞式飞机时代、喷气式飞机时代和信息时代 3 个阶段。	航空	军事

表 7　ChatGPT 错误分类案例

数目	图　　书	类目	
		真实	预测
1	标题：文明的力量 关键词：世界经济形势 摘要：全书以“讲好中国故事”为出发点，主要通过对国际时事风云的刻画和描述，客观、理性地记录了全球化与“逆全球化”背景下中国社会经济的各个方面和国际形势。	政法	经济
2	标题：中国发展简报 关键词：回望，纪实 摘要：本卷精选《中国发展简报》自 2001 年创刊到 2011 年十年间发表的优秀文章以飨读者。	经济	历史
3	标题：编剧原理 关键词：电影编剧 摘要：本书以“故事原理”为基础，分析了“编剧原理”的深层规律和时间技术。	文学	艺术
4	标题：发展哲学引论 关键词：社会发展，哲学理论 摘要：无。	历史	哲学

MAPL 大部分误分类都是源自标题与关键词信息误导。第 1、3、4 条案例表现出图书标题与已存在的类目重

合，例如《璀璨文化：中国音乐的发展历史与现代传承》体现出“音乐”，《航空技术与空中作战》出现“空战”等误导性信息。MAPL 研究方案对相关性类目出现误分类判断，例如“历史”与“政治”，“哲学”与“政治”，“军事”与“航空”等。我们分析其误差原因是目前预训练语言模型尚未达到人类语言理解水平，对类似词汇辨别能力不够。这一点将会随着自然语言处理、语言模型的发展而改善。

本研究无法限制 ChatGPT 预测类目范围，即其有可能超出本研究采集的类目范围。从表 7 可以观察到，ChatGPT 容易受到关键词的干扰而生成错误类目。第 1、2、3、4 条案例均是因为关键词存在误导性信息，例如“世界经济形势”指向“经济”类目，“电影编剧”指向“艺术”类目等，这说明当使用 ChatGPT 作为分类工具时，应当重点关注关键词信息。此外，第 4 条案例，ChatGPT 根据关键词信息，生成“哲学”类目，但该类目并不存在于样本集，这说明 ChatGPT 并不会完全按照本研究划定的范围生成答案。

此外，本研究将全部误分类图书的关键字、简介、标题等内容进行词云可视化，旨在直观地辨别研究方案在哪一类任务上频繁出现分类失误，起到一图胜千言的效果，如图 7 所示。从图中可以观察到，文化、政治、哲学在词云中出现频率最高，这表明研究方案对于文化、政治、哲学类书籍分类性能尚存在提升空间。我们分析其原因是这三类书籍本质上具有较强的跨学科性，经常在不同类型书籍中出现，对研究方案造成误导。同样，图 7 也可以表明，研究方案在音乐、地理、数学等其他类型书籍分类上表现出高准确性。该结果能够启迪图书馆员针对性使用研究方案，避免将其用于高误差场景。

图 7　MAPL 模型误分类词云图

4　应用前景研究

随着深度学习的不断发展，专业型智能工具已经开始应用到各行各业，例如同声传译、个性化推荐等。基于深度学习的图书分类技术可以提升分类效率、降低人力成本等。由于深度学习本身缺乏坚实的数学证明，基于深度学习的图书分类技术也不具备可解释性。因此，该方法适用于对分类精度与可解释性具有容忍性的应用场景。根据目前已取得的分类、预测准确率，有潜力充当人工标注的辅助工具。特别是提示学习在时效性、简洁性、小样本学习方面拥有传统深度学习方案无法比拟的优势，适用于要求短时间完成分类任务的场景。

本研究从郑州图书馆随机选择 300 本图书均分为互不重叠的三组，例如 A 组、B 组与 C 组。本研究将 A 组输入多模态注意力提示学习 MAPL 模型中获得最高预测类目，然后让 3 位志愿者在此结果上进行最终标注，他们能够在 0. 3 个小时完成 A 组 300 本书标注，其中 6 本书标注错误，准确率达到 98%。B 组不采用任何人工智能辅助工具，由 3 位志愿者直接标注，经过 1. 2 个小时才能够完成标注，其中 4 本书标注错误，准确率达 98. 6%。C 组输入多模态注意力提示学习模型中输出前 5 个预测类目，然后让 3 位志愿者依据 5 个预测类目进行最终标注，他们能够在 0. 4 个小时完成标注，其中 2 本书标注错误，准确率达到 99. 3%。

本研究采用同样的方式，将 A 组输入 ChatGPT 模型中获得最高预测类目，然后让 3 位志愿者在此结果上进行最终标注，他们能够在 0. 1 个小时完成 A 组 300 本书标注，其中 3 本书标注错误，准确率达到 99%。本研究将 C 组输入 ChatGPT 模型中输出前 3 个预测类目，然后让 3 位志愿者依据 3 个预测类目进行最终标引，他们能够在

0.2 个小时完成标引，其中 2 本书标注错误，准确率达到 99.3%。这表明 ChatGPT 可以辅助图书馆员从事图书分类任务。

5 结语

学习图书类目的特征与规律，准确高效地达到自动分类图书类目，避免高昂的人力成本，为开展图书数字化、智能化等工作铺垫基础，已经成为当前图书馆学的热门问题。已有的深度学习方法存在诸多问题，例如过度依赖大规模数据集、模型训练时间过长、无法适用于小样本图书分类场景等。为了解决该问题，本研究以广州图书馆与郑州图书馆的大规模图书为研究对象，将图书分类转换为完形填空，设计“询问大语言模型－提示－生成”的图书分类范式。本研究提出一种基于多模态注意力提示学习的图书自动分类方案，并具体实施两种路径。MAPL 通过预训练语言模型 BERT 与 ResNet 提取图书多模态特征，设计 5 种典型提示模板，采用注意力机制融合这些提示模板，将该模板直接输入预训练语言模型获得类目。ChatGPT 设计 5 种典型提示模板，直接输入 ChatGPT 模型即可获得类目。本研究在广州图书馆与郑州图书馆 10 个一级类目 114 823 条图书数据集上进行实验验证，在精准率、召回率与 F1 等指标上获得最优分类结果，验证本方案的可行性与有效性。

同时，本研究也存在一定的局限性。笔者仍采用硬提示的方式，即凭借专业知识设计出提示模板，而不同的提示模板对方案影响较大。因此，后续改进方案将考虑采用软提示方式，即避免人工设计提示模板。

参考文献

[1]于梦月，申静．基于知识融合的智库知识服务创新机制模型构建[J]．图书馆学研究，2023（9）：62－69.
[2]戎璐，张亚洲．一种注意力序列到序列模型的生成式层次文档分类[J]．图书馆学研究，2022（5）：45－56.
[3]Zhang Y，Wang J，Liu Y，et al. A Multitask Learning Model for Multimodal Sarcasm，Sentiment and Emotion Recognition in Conversations[J]．Information Fusion，2023（93）：282－301.
[4]Zhang M，Li J. A Commentary of GPT－3 in MIT Technology Review 2021[J]．Fundamental Research，2021（6）：831－833.
[5]张坤，王文韬，谢阳群．机器学习在图书情报领域的应用研究[J]．图书馆学研究，2018（1）：47－52.
[6]刘高军，陈强强．基于极限学习机和混合特征的中文书目自动分类模型研究[J]．北方工业大学学报，2018（5）：99－104.
[7]赵萌．基于增量学习的图书文本分类方法[J]．情报探索，2021（7）：52－56.
[8]Guo Shiqi，Yun Qiang，Chen Liang，et al. Research on the Method of Judging Reference Document in Patent Invalidation Using GBDT[J]．Library and Information Service，2021（2）：117－121.
[9]Safae L，El Habib B，Abderrahim T. A Review of Machine Learning Algorithms for Web Page Classification[C]//IEEE. In2018 IEEE 5th International Congress on Information Science and Technology（CiSt）. Marrakech：IEEE Digital Libraries，2018：220－226.
[10]Shah K，Patel H，Sanghvi D，et al. A Comparative Analysis of Logistic Regression，Random Forest and KNN Models for The Text Classification[J]．Augmented Human Research，2020（5）：1－6.
[11]Kim Y. Convolutional Neural Networks for Sentence Classification[C]//Association for Computational Linguistics. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing（EMNLP）. Doha：Association for Computational Linguistics，2014：1746－1751.
[12]郭利敏．基于卷积神经网络的文献自动分类研究[J]．图书与情报，2017（6）：96－103.
[13]吴洁，桂亮，刘鹏．基于图卷积网络的高质量专利自动识别方案研究[J]．情报杂志，2022（1）：88－95＋124.
[14]Zhang Yazhou，Liu Yaochen，Li Qiuchi，et al. CFN：A Complex－Valued Fuzzy Network for Sarcasm Detection in Conversations[J]．IEEE Transactions on Fuzzy Systems，2021（12）：3696－3710.
[15]Zhang Y，Song D，Zhang P，et al. A Quantum－Inspired Sentiment Representation Model for Twitter Sentiment Analysis[J]．Applied Intelligence，2019（49）：3093－3108.

戎　璐　郑州轻工业大学馆员，硕士。研究方向：图书咨询索引、图书自动化建设。