● 邱昕鹏，李　晶（中山大学信息管理学院，广东　广州　510006）

基于大语言模型的科技论文语义新颖性测度研究*

摘　要：［目的/意义］基于大语言模型实现科技论文问题词、方法词等关键词的半监督生成，并将生成的关键词应用于科技论文语义新颖性测度。［方法/过程］提出一种基于大语言模型的科技论文语义新颖性测度模型，通过LoRA和提示词协同微调以提升大语言模型的关键词生成和结构化测度准确性和区分度。［结果/结论］实验结果显示，模型生成的关键词召回率、精确率和调和平均数F_1达到62.5%、73.3%和67.5%的水平，召回率等指标随着训练样本增加而提高，但增长率存在先增后减趋势，训练集在3000时效果高且成本低。结果表明利用LoRA和提示词微调后的大语言模型具备较高性能，且所提出的科技论文语义新颖性测度方法具有有效性和鲁棒性。

关键词：科技论文评价；创新性测度；语义新颖性；大语言模型

DOI：10.16353/j.cnki.1000-7490.2025.06.022

引用格式：邱昕鹏，李晶.基于大语言模型的科技论文语义新颖性测度研究［J］.情报理论与实践，2025，48（6）：187-194.

Research on Scientific and Technological Paper Semantic Novelty Measurement Based on Large Language Model

Qiu Xinpeng，Li Jing

（School of Information Management，Sun Yat-sen University，Guangdong Guangzhou 510006）

Abstract：［Purpose/significance］Based on the large language model，the semi-supervised generation of keywords such as problem words and method words is implemented，and the generated keywords are applied to the semantic novelty measurement of scientific and technological papers.［Method/process］In this paper，a semantic novelty measurement model based on large language model is proposed. LoRA and prompt words are co-fine-tuned to improve the accuracy and discrimination of keyword generation and structural measurement of large language model.［Result/conclusion］The keyword recall rate，precision rate and F_1 generated by the model in this paper are 62.5%，73.3% and 67.5%. The recall rate and other indicators increase with the increase of training samples，but the growth rate tends to increase first and then decrease，and the training set has high effect and low cost at 3000.The experimental results show that the large language model after the collaborative fine-tuning of LoRA and prompt words has better performance，and the semantic novelty measurement method proposed in this paper is effective and robust.

Keywords：evaluation of scientific and technological papers；innovative measurement；semantic novelty；large language model

0　引言

2018年7月，中共中央办公厅和国务院办公厅联合印发《关于深化项目评审、人才评价、机构评估改革的意见》，要求“注重标志性成果的质量、贡献、影响”，2020年2月科技部、教育部也相继出台政策，包括《关于破除科技评价中“唯论文”不良导向的若干措施（试行）》《关于规范高等学校SCI论文相关指标使用　树立正确评价导向的若干意见》，系列政策直指科技评价要回归成果学术质量的本质，突出新颖性和贡献等特征的评价导向[1]。

科技论文是科研成果的主要载体形式，对科技论文进行评价有助于激发研究人员积极性，推动高质量的学术成果产出。目前针对科技论文新颖性的测度多以文献计量指标为主，包括科技论文的被引频次、合著特征等，但是总体来看，这些指标主要是通过科技论文的外部特征间接反映其质量，缺乏深入科技论文内容层面的测度方法[2]。

针对语义内容新颖性的测度逐渐受到研究者关注[3]。一方面，情报学、科学计量学逐步将研究对象从科技论文外部引用关系推进至科技论文微观语义知识单元[4]；另一方面，由于科技论文的外部引用关系一般需要较长时间才能体现出来，学者们开始使用语义新颖性来描述科技论文的创新特质[5]。然而，随着文本数据的日益增长，直接通过人工对海量科技论文或者其摘要进行语义层面新颖性测度变得效率低下。大语言模型（以下简称大模型）等日益成熟的人工智能技术为自动文本理解提供了技术支持，为科技论文的创新测量带来了发展机遇。因此，利用人工智能技术帮助科研人员从海量科技论文中选出具有语义新颖性的知识元，以节省科研人员的查找、选择和利用的时间，是科技论文新颖性评价的直接目的和重要价值[6]。在

* 本文为国家社会科学基金项目“科技论文创新质量的微观测度及应用研究”的成果，项目编号：22BTQ097。

测度方法层面，当今有不少学者基于深度学习和大模型等人工智能方法帮助人类进行自动关键词生成，但是仍然存在测度和识别结果区分度不高、准确性较弱以及人工标注成本过高等问题[7]。

相较于以往研究，本文着力于以下两个方面的创新研究。首先，本文使用高效参数LoRA微调训练大模型LLaMA3，同时基于上下文学习原理设计科技论文关键词生成的提示词，以提升模型文本处理效率。其次，将微调训练后的大模型LLaMA3应用于科技论文语义新颖性测度，力求提高结果区分度和准确性。本文的研究成果将为完善科技成果评价提供理论模型和测度方法，为开展面向不同类型科技成果语义新颖性创新评价提供量化指标设计和定量方法的支持。

1 相关研究

1.1 科技论文语义新颖性内涵

针对科技论文语义新颖性概念中的“科技论文语义内容”不少学者给出了阐述。黄红等[8]认为科技论文语义内容单元是文中表示特定科研要素的文本单元，通常包含研究背景、目的、问题、方法与结论等一系列要素，这些要素主要以词语、句子等文本形式表达。科技论文新颖性测度体现的是科技论文所提出的新理论、新方法、新论点或新结论等是否具有与已有研究不同的内容或研究范式[9]，通过对“科技论文语义内容”的“查新”能在一定程度上反映科技论文的语义新颖性[10]。因此，本文认为科技论文语义新颖性是指在科技论文中提出了不同于以往研究的语义内容，包括提出新的研究问题与方法、定义新概念、使用新研究技术与数据等，或者对已有研究语义内容的改进和完善，下文的模型设计和实证均围绕这一概念开展研究。

1.2 基于关键词的科技论文语义新颖性测度

在科技论文数量爆炸式增长背景下，许多研究者从文本关键词角度进行语义新颖性测度，认为未出现的或出现次数极少的关键词可能代表提出了新知识或新概念[11]。在测度方法层面，有学者使用传统统计方法对语义新颖性进行测度。例如，任海英等[12]使用主题词共线网络测度论文语义新颖性；沈阳[13]通过抽取关键词，基于关键词在文中出现频率、被用户检索频率等指标构建学术成果语义新颖性测度指标。也有学者使用机器学习和深度学习模型进行测度。例如，钱佳佳等[14]从词汇功能角度，将关键词分为问题词和方法词，基于BERT模型计算各种关键词的词共现率对论文语义新颖性进行测度；逯万辉等[15]基于HMM和Doc2Vec算法抽取论文主题词和关键词，从文本相似度角度计算论文语义新颖性。总体来看，基于关键词的生成和分析是科技论文语义新颖性测度的主要方法，但现有研究普遍存在两个问题：一是生成关键词的覆盖度问题，即模型训练集的缺失或不严谨可能会导致模型所生成的关键词准确性不高，导致测度区分度较差；二是结果可信度和区分度的问题，即科技论文语义新颖性测度结果与同行评议等“金指标”相比可信度如何，是否实现了区分度等效果上的优化。

因此，本文提出一种基于大模型微调技术的关键词生成方法，通过提示词模板输入LLaMA3大模型，再通过LoRA参数微调提高LLaMA3大模型的多头自注意力性能，加强模型从不同子空间中捕捉多种关键特征的能力[16]，提高LLaMA3的自动标注和自动生成性能，力求在生成更细粒度关键词的基础之上，提高科技论文语义新颖性测度的区分度和准确性。

2 研究方法与技术路线

本文基于大模型构建科技论文语义新颖性测度模型，选择总参数量为7 Billion的预训练LLaMA3模型作为基线大模型，通过微调技术训练大模型，得到适用于本文任务场景的模型，研究框架见图1，图中关键步骤将分小节论述。

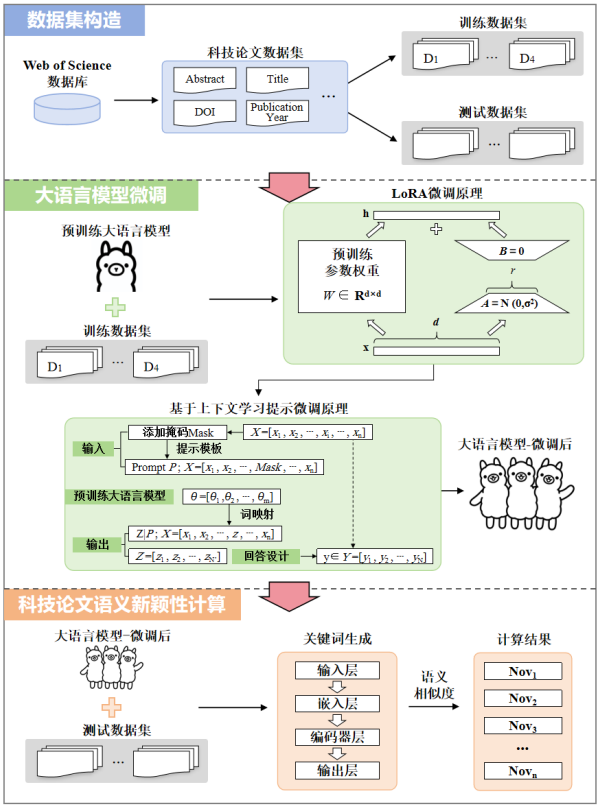

图1 基于大模型的科技论文语义新颖性测度研究框架

Fig. 1 Research framework of scientific and technological paper semantic novelty measurement based on large language model

2.1 基于LoRA的高效参数微调

为了解决大模型下游任务处理精度不够的问题，本文

针对大模型的超参数进行微调。具体而言，本文采用内在秩适配器LoRA框架，原理过程见图1。LoRA对大模型进行微调和训练，对预训练模型的权重矩阵$W_0 = R^{d\times d}$，通过低秩分解来表示其更新[17]，见公式（1）：

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

式中，W_0表示预训练模型的权重矩阵；ΔW表示微调时的参数更新；B是一个可训练的矩阵，其初始化时全为0矩阵。A也是一个可训练的矩阵，$r << \min(d, k)$，$B\in R^{d\times r}$，$A\in R^{r\times k}$，训练过程中，W_0不再进行梯度更新，A和B是可训练参数。输入x，模型的前向传播过程被更新，见公式（2）：

$$h = W_0x + \Delta Wx = W_0x + BAx \tag{2}$$

式中，W_0x表示原始的前向传播过程，即输入x通过权重矩阵W_0得到输出。

2.2 基于上下文学习(ICL)的提示词微调

为了解决语义差异问题和过拟合问题，提高大模型在特定任务上的适应性，本文在LoRA高效参数微调后，基于ICL原理对大模型进行提示微调，其作用主要是从科技论文摘要中准确识别并生成与原文相关的关键词，主要原理见图1。ICL的本质思想是类比学习[18]，这是一种离散调优的方法，该方法着力于调整模型的提示或标记，而保持底层语言模型参数不变，以提高模型的性能，所需的计算资源和存储空间也大大减少[19]，模板的构成形式见公式（3）：

$$\text{Prompt} = \text{Instruction} + \text{Input} + \text{Example} + \text{Output} \tag{3}$$

式中，Instruction表示对关键词生成任务的说明，力求清晰、具体地描述任务目标和任务内容；Input表示输入的文本；Example表示本文根据任务给出的示例；Output表示输出结果要求，设定为生成结果控制在6个关键词。本文根据ICL内涵和提示词工程模板，设计提示词见图2。

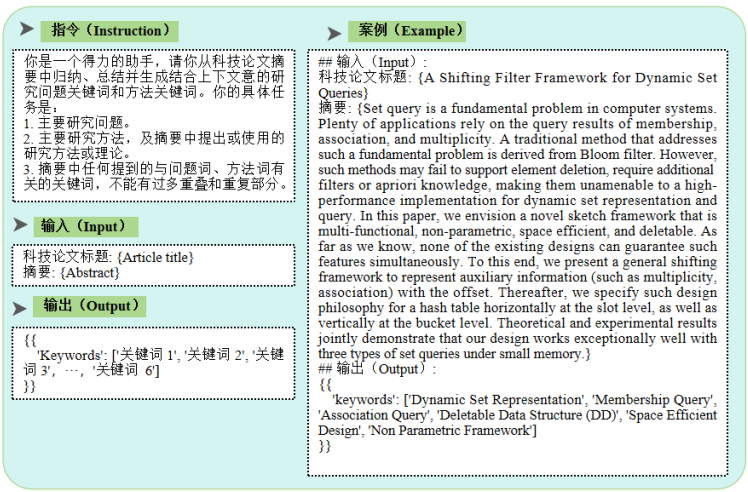

图2 基于ICL的提示词设计框架

Fig. 2 ICL-based prompt words design framework

2.3 科技论文语义新颖性算法

本文将生成的每篇样本论文的关键词汇总，以样本论文的发表年份为参照点，使用微调后大模型LLaMA3进行语义相似度计算。具体地，对比在相同研究领域内、早于当前样本论文发表时间的其他研究文献中的关键词相似度，将这些关键词出现频次的数据进行记录，并标记为特定的参照数据。记作$n(Q_k)$。代入计算公式，所得分数即为样本论文语义新颖性分数[14]。

本模型引用关键词语义新颖性测度公式，测度见公式（4）：

$$\text{Nov}_n = \frac{\sum_{k=1}^{|Q|}\frac{1}{\ln\left[n(Q_k)+1\right]+1}}{|Q|} \tag{4}$$

式中，Nov_n表示样本论文的语义新颖性分数；Q表示样本论文关键词集合；$|Q|$表示集合Q的元素个数；Q_k表示样本论文关键词相较于其发表前论文中出现的第k个词；$n(Q_k)$表示样本论文关键词相较于其发表前论文中出现频次。

3 实证分析

3.1 样本数据与技术参数

本文以Web of science核心集作为数据源，选取计算机科学一级学科2022—2023年度发表的科技论文作为实验样本，经检索得到样本论文共27078篇。鉴于科技论文语义新颖性的评估需将文献内容与其发表之前的同领域作品进行对比，经检索得到全库数据科技论文共205967篇。所有数据信息包括论文题目、摘要、被引频次、发表期刊JCR分区等关键信息，旨在对这一特定时段内科技论文语义新颖性进行实证分析。将针对本文特定任务需求的样本数据集分为训练集和测试集，其中训练集用于训练大模型。本文选取该学科2022年的数据集共12541篇作为训练集，构建了样本数量从小到大4个训练数据集D_1、D_2、D_3、D_4，其中，D_1、D_2、D_3分别包含500、1500、3000个样本，D_4包含2022年全样本数据共12541篇；在测试集选取上，选取该学科中2023年的数据集共14537篇作为测试集，使用测试集来评估模型的最终性能。

实验环境所使用的GPU为NVIDIA A800 SXM4，Python版本为3.10，使用的Pytorch版本为2.0.1。本文大模型训练的超参数设置见表1。

3.2 样本数据的训练效果评价

文本关键词抽取效果评价中，精确率、召回率和调和平均数F_1值作为抽取效果评价指标被广泛使用[20]，计算公式见公式（5）～公式（7）：

表1 超参数设置

Tab. 1 Parameter setting

超参数类型	数值
学习率(learning_rate)	2e-5
训练批次大小(per_device_train_batch_size)	4
评估批次大小(per_device_eval_batch_size)	4
梯度累积步骤(gradient_accumulation_steps)	8
训练轮次(num_train_epochs)	3
权重衰减(weight_decay)	0.01
评估策略(evaluation_strategy)	'steps'
评估间隔(eval_steps)	500
保存间隔(save_steps)	500

$$P = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \times 100\% \quad (5)$$

$$R = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \times 100\% \quad (6)$$

$$F_1\text{值} = \frac{P \times R \times 2}{(\mathrm{P} + \mathrm{R})} \times 100\% \quad (7)$$

式中，TP表示预测准确样本中正例数；FP表示预测为正但实际为负的样本个数；FN表示预测为负例但实际为正例样本个数；P表示精确率；R表示召回率；F_1值表示调和平均数。

3.2.1 训练样本数对结果的影响 首先基于D_1、D_2、D_3、$D_4$4组不同的训练集分别对LLaMA3大模型进行微调，将4组数据训练出的模型运用于测试集的关键词生成任务当中，并分别计算4组结果的精确率、召回率和F_1值，结果见表2。

表2 基于不同训练样本数微调后LLaMA3关键词生成效果比较

Tab. 2 Comparison of keywords generation effect of LLaMA3 after fine-tuning based on different training sample numbers

数据集	召回率(%)	精确率(%)	F_1值(%)
D_1	33.5	54.0	48.5
D_2	44.5	68.0	53.8
D_3	62.5	73.3	67.5
D_4	**64.0**	**75.5**	**69.3**

注：粗体表示最优值。

由表2结果可知，从D_1训练集到D_4训练集的数据量增加，模型召回率从33.5%上升到64.0%，精确率由54.0%上升到75.5%，F_1值更是由48.5%上升为69.3%，三个评价指标的涨幅较明显，因此随着训练样本数的增加，测试集的关键词生成效果也增加。但是观察到不同数据组之间的效果涨幅存在先增后减的趋势，以F_1值为例，D_1训练集与D_2训练集之间的F_1值增长5.3%，D_2训练集与D_3训练集之间的F_1值差为13.7%，三组训练集之间的涨幅明显增加，然而D_3训练集与D_4训练集之间的F_1值差值降为1.8%，涨幅明显减少，但D_4训练集所需要的硬件要求和时间远高于D_3训练集，因此，出于硬件环境成本方面的考虑，后续实验均采用D_3训练集进行大模型训练。

3.2.2 与通用大模型的比较分析 根据上述实验结果，本文选用D_3作为训练集。为验证本文微调模型的关键词生成效果，选取目前具有代表性的通用大模型Gemma2、Phi3、GPT-4、LLaMA3与本文微调模型进行比较，结果见表3。

表3 微调大模型与通用大模型的关键词生成效果比较

Tab. 3 Comparison of keyword generation effect between fine-tuned large language model and general large language model

模型	召回率(%)	精确率(%)	F_1值(%)
Phi3	46.0	44.0	47.0
Gemma2	38.0	54.5	44.8
LLaMA3	44.5	56.0	49.6
GPT-4	56.4	60.3	58.3
本文模型	**62.5**	**73.3**	**67.5**

注：粗体表示最优值，出于控制成本考虑，后续研究使用D_3训练集进行训练。

表3展示了基于D_3训练集微调的大模型和通用大模型Gemma2、Phi3、GPT-4、LLaMA3的关键词生成效果比较。在通用大模型中，闭源的GPT-4的抽词效果最佳，F_1值高达58.3%，召回率和精确率也都远远超过其他开源大模型。在开源基线大模型中，LLaMA3的F_1值为47.0%，Gemma2和Phi3的关键词生成效果相近，F_1值处在44.8%和49.6%左右的水平，处于较弱的水平。相较于GPT-4，开源的LLaMA3模型规模较小，经过微调后，LLaMA3关键词生成效果明显提高，显著优于其他通用大模型，精确率高达73.3%，F_1值达到67.5%的水平，超过GPT-4。因此，相关数据表明本文模型能够识别更加符合上下文意的语义关键词，实现了更精准、细致的科技论文语义识别，为后续科技论文语义新颖性的测度增加可信度。

3.2.3 消融实验 为了验证本文针对科技论文语义新颖性测度模型的两步微调对关键词生成的有效性，同时据上述实验结果，采用D_3训练集来进行消融实验。具体地，将本文微调模型与仅有LoRA微调的LLaMA3、仅有提示词微调的LLaMA3、基线模型LLaMA3的测试集效果进行对比分析，结果见表4。

相比于本文模型，提示词微调技术使得LLaMA3大模型的表现有所提升，召回率和F_1值分别上升了了10.5%和9.5%，这表明在LLaMA3模型中整合更精细、规范的提示模板，能够促进大模型聚焦于任务的先验知识，进而确保所构建的模型内嵌有科技论文摘要的上下文语义信息，使模型表示得以更精准地匹配下游任务的需求[21]。相比于本文模型，LoRA微调技术也具有一定模型优化效果，召回率和F_1值分别上升了16.5%和13.1%，性能得到了提升。综合消融实验结果来看，本文所使用的提示微调和LoRA

表4　消融实验结果

Tab. 4　Results of ablation experiments

模型	召回率（%）	精确率（%）	F_1值（%）
LLaMA3基线模型	44.5	56.0	49.6
LLaMA3-提示词微调	55.0	65.5	59.8
LLaMA3-LoRA参数微调	61.0	63.4	62.7
本文模型	**62.5**	**73.3**	**67.5**

注：粗体表示最优值，出于控制成本考虑，后续研究使用D_3训练集进行训练。

微调技术均对相关任务起到了显著的优化作用。

3.3　科技论文语义新颖性测度结果

本节基于本文所提出的微调LLaMA3大模型与其他通用大模型Gemma2、Phi3、GPT-4、LLaMA3基线模型作为工具测度出的样本论文语义新颖性分数，通过区间分布图来反映不同模型计算出的结果，将所得到的结果数据进行区间分布统计，按区间将科技论文语义新颖性取值分为5个部分，结果见图3。

整体来看，5个模型测量结果的中位数大致都略微低于其平均数，这表明样本中大部分的语义新颖性得分处于较高层次，同时也反映出计算机科学领域科技论文的研究关键词更新速度相对较快。与同为开源模型的Gemma2、Phi3对比来看，本文模型测度结果显示，分数为1的科技论文有2633篇，占比为18.1%，而Gemma2模型和Phi3模型的结果显示分数为1的科技论文分别占比30.2%和29.0%，同时本文模型各区间具有更平均的分布。与GPT-4大模型相比，微调后的LLaMA3大模型的测度结果区间显示要比GPT-4具有更宽泛的范围，GPT-4对科技论文语义新颖性分数较低的样本区分度较差。本文模型测度结果极值明显较大，并且四分位距明显大于其余模型的结果，反映微调后的LLaMA3模型测度的科技论文语义新颖性分数分布更加分散，具有更高的科技论文语义新颖性区分效果。

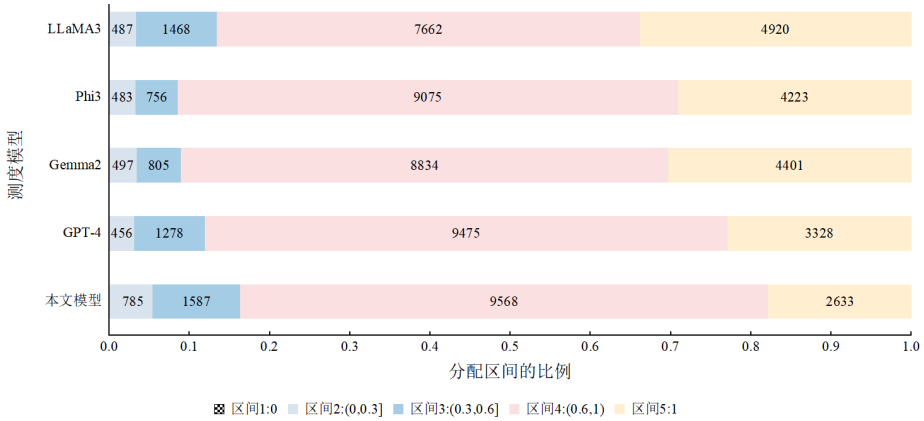

图3　基于不同模型的科技论文语义新颖性分数分布区间图

Fig. 3　Interval plot of scientific and technological paper semantic novelty score distribution based on different models

3.4　有效性检验

3.4.1　不同语义新颖性测度模型对比分析　为了研究本文方法与现有方法之间的异同，选取一种基于融合大小语言模型LLaMA3-BPE的科技论文语义新颖性模型作为参照[22]进行对比分析，具体分布可见图4。在对比分析中，LLaMA3-BPE模型的结果展现出了较为明显的两极分化趋势，其计算所得的分值主要密集分布在[0.7，0.8)区间。相比之下，本文所提出的语义新颖性测度模型在结果分布上展现出了更为均衡的特点。表明在面对大规模数据集时，仅仅依靠大小模型的简单结合所实现的协同效果，在处理上下文信息方面可能效果一般。本文所提出的方法不仅增强了模型捕捉细微差异的能力，还进一步提升了科技论文语义新颖性测度的区分度和准确性。

3.4.2　基于德尔菲法的结果有效性检验　为了进一步确保模型测度的有效性，本文参照段美珍等[23]的研究方法，使用德尔菲法，抽取科技论文语义新颖性分数排名较高的10篇样本论文和排名较低的10篇样本论文，将排名打乱后，邀请13名中山大学计算机学院的专家对这20篇样本论文进行调查，其中2位具有高级职称，4位是博士研究生，7位是硕士研究生。关于样本论文新颖性的评分规则，让专家按照该篇科技论文的语义新颖性“非常好、好、一般、较差、非常差”5个等级分别赋分5、4、3、2、1。13名专家打分后，汇总数据并计算所有专家的加权平均分，与本模型实证结果进行对比分析。经过对专家意见的收集和统计计算，

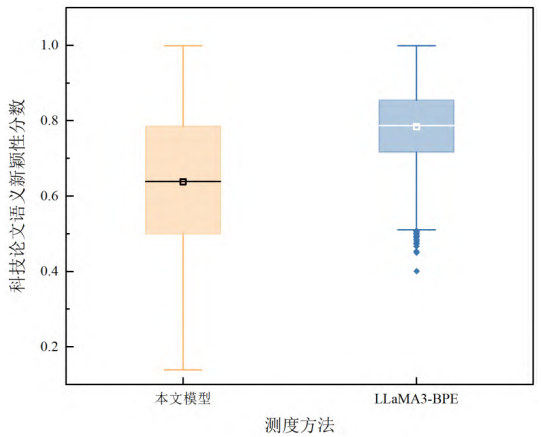

图4　两种科技论文语义新颖性测度方法比较

Fig. 4　Comparison of two methods for measuring the semantic novelty of scientific and technological papers

调查的响应率为100%，确保了调查的科学性和可信度。

结果显示，协调系数<0.7，且协调系数对应的P值均小于0.05，V_i均小于0.25，表明结果具有较高的协调性和一致性。表5显示，S_1组中，被专家打出平均分3～5分的论文占80%，打出低分的占比仅仅10%，说明本文模型能够一定程度上筛选出具有较高语义新颖性的论文。在S_2组中，被专家打出平均分3～5分的论文仅仅占20%，打出1～2低分的样本占比则达到70%，说明本文模型能够有效测度出语义新颖性较弱的论文，结果具有一定可靠性。

进一步地，在表6中，选取序号1和2的科技论文进行深入语义层面的分析，这两篇科技论文的语义新颖性分数均为最高分1分，专家组分别打出了4.462和4.384的高分。其中，题目为“Interval Dominance-Based Feature Selection for Interval-Valued Ordered Data”的论文发表在“IEEE Transactions on Neural Networks and Learning Systems”期刊上，该期刊位于JCR的Q1区。专家指出文章创新在于提出区间优势度与重叠度新阈值，并据此研究区间值优势粗糙集方法（IV-DRSA）及其性质，建立相关特征选择规则与算法。这些创新丰富了粗糙集理论应用，并为多值有序数据特征选择提供新工具。另一篇文章题目为“Regret-Theoretic Multi-attribute Decision-Making Model Using Three-Way Framework in Multi-scale Information Systems”，发表在“IEEE Transactions on Cybernetics”期刊上，该期刊位于JCR的Q1区。专家指文章创新融合3WD模型与后悔理论，构建多尺度信息系统多属性决策的“后悔—喜悦超序关系”，实现三域划分与相对接近度排序，解决误分类，反映决策者风险态度，具有高新颖性。结果显示，本文模型在确保信度基础上，能进行更细粒度、更精准的科技论文语义新颖性测度，为客观量化评价提供新理论与方法。

表5 专家评分结果汇总

Tab. 5 Summary of experts scoring results

组别	平均分位于1～2之间	占比(%)	平均分位于2～3之间	占比(%)	平均分位于3～5之间	占比(%)
S_1	1	10	1	10	8	80
S_2	7	70	1	10	2	20

注：S_1表示科技论文语义新颖性分数较高的10篇论文组；S_2表示科技论文语义新颖性分数较低的10篇论文组。“平均分位于3～5之间”表示该组科技论文中获得专家平均分位于3～5分的数量，“占比”表示该组论文中平均分位于特定区间的占比。

表6 德尔菲法调查结果(部分)

Tab. 6 Results of the Delphi method survey (part)

序号	科技论文标题	Nov	N_i	k_i	V_i
1	Interval Dominance-Based Feature Selection for Interval-Valued Ordered Data	1.000	4.462	0.692	0.197
2	Regret-Theoretic Multi-attribute Decision-Making Model Using Three-Way Framework in Multi-scale Information Systems	1.000	4.384	0.769	0.198
3	Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm	1.000	4.462	0.846	0.174
4	Adaptive Bipartite Output Tracking Consensus in Switching Networks of Heterogeneous Linear Multiagent Systems Based on Edge Events	1.000	4.308	0.846	0.174
5	MGRCDA: Metagraph Recommendation Method for Predicting CircRNA-Disease Association	1.000	4.154	0.923	0.134
6	Convolutional neural network based object detection system for video surveillance application	0.158	3.846	0.769	0.144
7	Federated deep learning for anomaly detection in the internet of things	0.155	1.846	0.000	0.203
8	An Adaptive Deep Learning Neural Network Model to Enhance Machine-Learning-Based Classifiers for Intrusion Detection in Smart Grids	0.152	1.769	0.000	0.248
9	Analysis of cutting-edge technologies for enterprise information system and management	0.148	1.769	0.000	0.248
10	CLAS: A new deep learning approach for sentiment analysis from Twitter data	0.145	1.846	0.846	0.203
...	...	...	...	...	...

注：序号1～5分别表示科技论文语义新颖性分数较高的5篇论文；序号6～10分别表示科技论文语义新颖性分数较低的5篇论文；Nov表示科技论文语义新颖性分数；N_i表示专家对序号i科技论文的平均分；k_i表示序号i科技论文被专家打出3～5分的频率；V_i表示各专家对序号i科技论文意见的均方差。

3.5 鲁棒性检验

为了检测模型的鲁棒性，本文选取其他开源大模型Phi3和Gemma2，沿用上文的D_3训练集进行微调训练。在经过LoRA和提示词微调后，对两个大模型输入相同测试集，结果见图5，微调后的Phi3得到的结果中，最大值为1，最小值接近0.1，而基线Phi3模型的结果显示最大值均为1，最小值处在0.45左右水平，Gemma2大模型经过微调后也呈现相似的趋势。然而Phi3和Gemma2两个基线大模型测度结果呈现较为集中趋势，科技论文语义新颖性区分度较差。因此，在经过本文模型相同的处理后，所得结果

区分度均比基线模型有了明显的提高，与LLaMA3大模型的测度结果之间差异很小，证明本文所提出的模型能够在一定程度上提高科技论文语义新颖性测度的区分度、准确性和科学性。

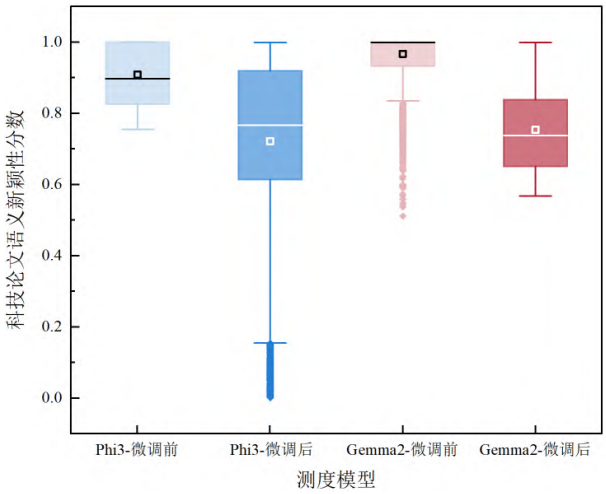

图5　微调前后的科技论文语义新颖性分数对比

Fig. 5　Comparison of semantic novelty scores of scientific and technological papers before and after fine-tuning

4　结果与讨论

本文提出基于微调训练LLaMA3大模型的科技论文语义新颖性测度的理论框架。实证研究中，本文以计算机一级学科2022年发表的科技论文作为训练集，2023年发表的科技论文为测试集。通过不同训练集数量结果，发现随训练集数量增加，召回率等指标的增长率呈先增后减趋势，可根据实际成本选择最优的训练量。在模型效果验证上，通过关键词生成效果评价验证了本文提出的模型能够实现语义内容关键词生成任务上的提升，通过消融实验验证了LoRA微调和提示词微调均会对LLaMA3大模型的关键词生成效果实现明显优化，为后续科技论文语义新颖性测度奠定基础。为了验证本文模型在科技论文语义新颖性测度中的信效度，针对测度结果展开了一系列对比分析，证明了其对科技论文语义新颖性测度任务有显著优化作用，提高了结果区分度和准确性。本文在最后章节通过替换通用大模型，检测了科技论文语义新颖性测度模型的鲁棒性和稳健性。在未来，可以进一步将本文大模型微调思路应用于中文文本的训练中，以期能够实现对中文论文的语义新颖性进行测度。□

参考文献

[1] 为建设世界科技强国而奋斗——在全国科技创新大会、两院院士大会、中国科协第九次全国代表大会上的讲话［EB/OL］.（2016-05-30）［2024-10-27］.https://news.cnr.cn/native/gd/20160531/t20160531_522287749.shtml.

[2] 曾建勋.推动科研论文语义评价体系建设［J］.数字图书馆论坛，2021（11）：1.（ZENG Jianxun. Promote the construction of semantic evaluation system for scientific research papers［J］. Digital Library Forum，2021（11）：1.）

[3] 孙晓玲.科技论文写作旨在突出创新［J］.西北师范大学学报（自然科学版），2007，43（6）：111-114.（SUN Xiaoling. Highlighting new ideas is the core of writing scientific paper［J］. Journal of Northwest Normal University （Natural Science），2007，43（6）：111-114.）

[4] 姜春林，张立伟，谷丽，等.知识单元视角下学术论文评价研究［J］.情报杂志，2014，33（4）：29-34.（JIANG Chunlin，ZHANG Liwei，GU Li，et al. Research on the evaluation of academic papers from the perspective of knowledge unit［J］. Journal of Intelligence，2014，33（4）：29-34.）

[5] 逯万辉，苏金燕，余倩.学术成果主题新颖性与学术引用的相关关系研究［J］.情报资料工作，2018（6）：68-73.（LU Wanhui，SU Jinyan，YU Qian. Research on correlation of academic achievement theme novelty and academic citation［J］. Information and Documentation Services，2018（6）：68-73.）

[6] 索传军，于果鑫.学术论文研究亮点的语言学特征与分布规律研究［J］.图书情报工作，2020，64（9）：104-113.（SUO Chuanjun，YU Guoxin. Exploration of the research “Highlights” in academic papers［J］. Library and Information Service，2020，64（9）：104-113.）

[7] 戎军涛，索传军，周彦廷，等.基于创新知识元谱系的学术论文新颖性测度研究［J］.图书情报工作，2024，68（1）：27-38.（RONG Juntao，SUO Chuanjun，ZHOU Yanting，et al. Novelty measurement of academic literature based on the innovative knowledge elements genealogy［J］. Library and Information Service，2024，68（1）：27-38.）

[8] 黄红，陈翀，张婧莹.科技文献内容语义识别研究综述［J］.情报学报，2022，41（9）：991-1002.（HUANG Hong，CHEN Chong，ZHANG Jingying. Review on identifying the semantics of scientific literature content［J］. Journal of the China Society for Scientific and Technical Information，2022，41（9）：991-1002.）

[9] 周露阳.论审评学术论文创新因素的指标体系［J］.编辑学报，2006（1）：68-70.（ZHOU Luyang. Index system for identifying innovation factors in academic papers［J］. Acta Editologica，2006（1）：68-70.）

[10] 魏绪秋，申力旭.学术论文创新性研究述评［J］.图书情报知识，2022，39（4）：68-79.（WEI Xuqiu，SHEN Lixu. A research review of the academic paper innovativeness［J］. Documentation，Information & Knowledge，2022，39（4）：68-79.）

[11] UDDIN S，KHAN A. The impact of author-selected keywords on citation counts［J］. Journal of Informetrics，2016，10（4）：1166-1177.

[12] 任海英，王德营，王菲菲.主题词组合新颖性与论文学术影响力的关系研究［J］.图书情报工作，2017，61（9）：87-93.（REN Haiying，WANG Deying，WANG Feifei. Relationship between novelty of key-term combinations and papers’ scientific impact［J］. Library and Information Service，2017，61（9）：87-93.）

[13] 沈阳.一种基于关键词的创新度评价方法[J].情报理论与实践，2007（1）：125-127.（SHEN Yang. A keyword-based innovation evaluation method [J]. Information Studies：Theory & Application，2007（1）：125-127.）

[14] 钱佳佳，罗卓然，陆伟.基于问题—方法组合的科技论文新颖性度量与创新类型识别[J].图书情报工作，2021，65（14）：82-89.（QIAN Jiajia，LUO Zhuoran，LU Wei. Novelty measurement and innovation type identification of scientific literature based on question-method combination [J]. Library and Information Service，2021，65（14）：82-89.）

[15] 逯万辉，谭宗颖.学术成果主题新颖性测度方法研究——基于Doc2Vec和HMM算法[J].数据分析与知识发现，2018，2（3）：22-29.（LU Wanhui，TAN Zongying. Measuring novelty of scholarly articles [J]. Data Analysis and Knowledge Discovery，2018，2（3）：22-29.）

[16] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems，Long Beach，CA，USA. New York：Curran Associates，2017：6000-6010.

[17] HU E J，SHEN Yelong，WALLIS P，et al. LoRA：low-rank adaptation of large language models [J/OL]. （2021-10-16）. [2024-10-27]. https：//doi.org/10.48550/arXiv.2106.09685.

[18] LUO Man，XU Xin，LIU Yue，et al. In-context learning with retrieved demonstrations for language models：a survey [J/OL]. （2024-03-23） [2024-10-27]. https：//doi.org/10.48550/arXiv.2401.11624.

[19] 高峰，刘晴，靳英辉，等.基于知识微调和信息融合的医学指南知识抽取[J/OL].武汉大学学报（理学版），1-11 [2024-10-27]. https：//doi.org/10.14188/j.1671-8836.2024.0032.（GAO Feng，LIU Qing，JIN Yinghui，et al. Medical guide knowledge extraction based on knowledge fine tuning and information fusion [J/OL]. Journal of Wuhan University（Natural Science Edition），1-11 [2024-10-27]. https：//doi.org/10.14188/j.1671-8836.2024.0032.）

[20] 沈思，陈猛，冯暑阳，等.ChpoBERT：面向中文政策文本的预训练模型[J].情报学报，2023，42（12）：1487-1497.（SHEN Si，CHEN Meng，FENG Shuyang，et al. ChpoBERT：a pre-trained model for Chinese policy texts [J]. Journal of the China Society for Scientific and Technical Information，2023，42（12）：1487-1497.）

[21] LIU Xiao，JI Kaixuan，FU Yicheng，et al. P-Tuning：prompt tuning can be comparable to fine-tuning across scales and tasks [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics （Volume 2：Short Papers）. Dublin，Ireland. Stroudsburg：Association for Computational Linguistics，2022：61-68.

[22] 李晶，邱昕鹏.基于大语言模型的论文创新质量测度研究[J].情报理论与实践，2025，48（3）：169-177.（LI Jing，QIU Xinpeng. Research on innovation quality measurement of papers based on large language model [J]. Information Studies：Theory & Application，2025，48（3）：169-177.）

[23] 段美珍，初景利，张冬荣，等.智慧图书馆建设评价指标体系构建与解析[J].图书情报工作，2021，65（14）：30-39.（DUAN Meizhen，CHU Jingli，ZHANG Dongrong，et al. Research on the construction of evaluation index system of smart library development in colleges [J]. Library and Information Service，2021，65（14）：30-39.）

作者简介：邱昕鹏，男，2000年生，硕士生。研究方向：智能评价与决策支持。**李晶**（通信作者，Email：lijing359@mail.sysu.edu.cn），女，1985年生，博士，副教授，博士生导师。研究方向：智能评价与决策支持，信息用户与信息治理。

录用日期：2025-01-09

（上接第177页）

[35] CAO Tianshi，LAW M T，FIDLER S. A theoretical analysis of the number of shots in few-shot learning [EB/OL]. [2024-12-20]. https：//arxiv.org/pdf/1909.11722.

[36] BROWN T B，MANN B，RYDER N，et al. Language models are few-shot learners [C]//Proceedings of the 34th International Conference on Neural Information Processing System. New York：ACM，2020：1877-1901.

[37] FULFORD I，NG A. ChatGPT prompt engineering for developers [EB/OL]. [2024-12-20]. https：//github.com/VeldiBharath/ChatGPT-Prompt-Engineering-for-Developers.

[38] 钱力，刘志博，胡懋地，等.AI就绪的科技情报数据资源建设模式研究[J].农业图书情报学报，2024，36（3）：32-45.（QIAN Li，LIU Zhibo，HU Maodi，et al. Construction model of AI-Ready for scientific and technological intelligence data resources [J]. Journal of Library and Information Science in Agriculture，2024，36（3）：32-45.）

作者简介：黄永文（ORCID：0000-0001-8912-0698），女，1975年生，博士，研究员。研究方向：科学数据管理，知识组织与知识服务。**马玮璐**，男，1999年生，博士生。研究方向：知识组织与知识服务。**鲜国建**，男，1982年生，博士，研究员，博士生导师。研究方向：关联数据与知识服务。**李娇**（ORCID：0000-0002-8876-3728），女，1989年生，博士，副研究员。研究方向：文本挖掘，知识图谱。**罗婷婷**，女，1985年生，硕士，副研究员。研究方向：知识组织，大数据融汇治理。**孙坦**（ORCID：0000-0002-8257-5064，通信作者，Email：suntan@caas.cn），男，1970年生，博士，研究员，博士生导师。研究方向：数字信息描述与组织。

作者贡献声明：黄永文，设计研究方案，实验设计与结果分析，撰写论文。**马玮璐**，数据处理及实验实现。**鲜国建**，数据采集，论文修改。**李娇**，文献调研与数据分类。**罗婷婷**，数据收集与数据分类。**孙坦**，提出写作思路，论文修改。

录用日期：2025-02-26