知网个人查重服务报告单(全文标明引文)

报告编号:BC2025073111136388500682611 检测时间:2025-07-31 11:36:38

篇名：大模型驱动的学术文本挖掘——推理端指令策略构建及能力评测

作者：陆伟;刘寅鹏;石湘;刘家伟;程齐凯;黄永;汪磊

检测类型：职称评审

比对截止日期：2024-08-23

检测结果

去除本人文献复制比：0% 去除引用文献复制比：0% 总文字复制比：0%

单篇最大文字复制比：0%

重复字符数：[0] 单篇最大重复字符数：[0] 总字符数：[15607]

0%(0) 0%(0) 大模型驱动的学术文本挖掘——推理端指令策略构建及能力评测_第1部分（总9706字）

0%(0) 0%(0) 大模型驱动的学术文本挖掘——推理端指令策略构建及能力评测_第2部分（总5901字）

（注释：无问题部分 文字复制部分 引用部分）

1. 大模型驱动的学术文本挖掘——推理端指令策略构建及能力评测_第1部分 总字符数：9706

相似文献列表

去除本人文献复制比：0%(0) 去除引用文献复制比：0%(0) 文字复制比：0%(0)

原文内容

大型语言模型突出的任务理解和指令遵循能力，使用户可以通过简单的指令交互完成复杂的信息处理任务。科技文献分析领域正在积极探索大模型的应用，但尚未形成对指令工程技术和模型能力边界的系统性研究。本文以学术文本挖掘任务为切入点，从上下文学习、思维链推理等角度设计推理端指令策略，构建了涵盖文本分类、信息抽取、文本推理和文本生成4个能力维度共6项任务的大模型学术文本挖掘专业能力评测框架，并选取了7个国内外主流的指令调优模型进行实验，对比了不同指令策略的适用范围和不同参数模型的专业能力。实验结果表明，少样本、思维链等复杂指令策略在分类任务上的应用效果并不显著，而在抽取、生成等难度较高的任务上表现良好。千亿级参数规模的大模型经过指令引导，能够取得与充分训练的深度学习模型相近的效果，但对于十亿级或百亿级规模大模型，推理端的指令策略存在明显上限。为了实现大模型向科技情报领域的深层嵌入，现阶段仍需在调优端对模型参数进行领域化适配。

Task comprehension and instruction-following abilities of large language models enable users to complete complex information-processing tasks through simple interactive instructions. Scientific literature analysts are actively exploring the application of large language models; however, a systematic study of the capability boundaries of large models has not yet been conducted. Focusing on academic text mining, this study designs inference-end prompting strategies and establishes a comprehensive evaluation framework for large language model-driven academic text mining, encompassing text classification, information extraction, text reasoning, and text generation, covering six tasks in total. Mainstream instruction-tuned models were selected for the experiments, to compare the different prompting strategies and professional capabilities of the models. The experiments indicate that complex instruction strategies, such as few-shot and chain-of-thought, are not effective in classification tasks, but perform well in more challenging tasks, such as extraction and generation, whereby trillion-parameter scale models achieve results comparable to those of fully trained deep-learning models. However, for models with billions or tens of billions of parameter scales, there is a clear upper limit to inference-end instruction strategies. Achieving deep integration of large language models into the field of scientific intelligence requires adaption of the model to the domain at the tuning end.

大模型;学术文本挖掘;指令工程;能力评测;large language model;academic text mining;instruction engineering;capability evaluation

0 引言

经过指令微调和人类意图对齐后的大规模语言模型（large language model,LLM）能够应对不同场景的自然语言处理任务

，影响着各个垂直领域的智能技术应用范式，包括以学术文本为主要研究对象的科技情报领域。科技情报工作的核心是从科技文献内容中挖掘和利用知识，从而实现科技文献表层信息组织到深层语义组织的跨越[1]。多年来，科技文献分析研究者运用自然语言处理、深度学习等人工智能技术，着力于解决引文推荐与功能识别[2]、论证区间自动分类[3]、科技信息抽取[4-5]等学术文本挖掘任务，取得了卓有成效的进展，研究成果被成功应用于科技情报整编与科技信息服务等业务场景[6]，但也存在模型性能依赖于大规模高质量标注数据，微调模型泛化能力差等诸多问题。大模型强大的少样本学习能力，使其仅依靠少量示例便能在各类下游任务中取得优异的性能；生成式的任务处理逻辑，打破了学术文本挖掘任务之间的界限。这些能力一定程度上解决了传统深度学习模型训练难、迁移难的痛点。

然而，大模型如何全方位嵌入科技情报领域，各类大模型在学术文本挖掘等领域任务上的专业能力如何，这些问题仍有待探究。为了探索大模型在科技情报领域的应用模式，提升大模型驱动的学术文本挖掘性能，本文设计了大模型驱动的学术文本挖掘框架，如图1所示。在该框架中，大模型应用于科技文本挖掘的策略可归纳为推理端的指令策略和调优端的微调策略两个方面。推理端是指在不调整模型参数的前提下，通过运用上下文学习、思维链推理等指令工程技术，构造自然语言指令充分挖掘大模型蕴含的知识，调动模型解决各类学术文本挖掘任务[7]；调优端则是运用LoRA（low-rank adap‐tation）、P-tuning v2等参数高效微调策略[8]，在节省计算资源的前提下，将大模型的通用能力适配为面向特定领域的专业能力。例如，在应对摘要句功能分类任务时，在推理端，可以人工编写或检索相似样本示例，构造指导模型进行逐步推理的思维链，通过将样本示例、思维链添加到指令模板中，最终组成模型上下文；在调优端，选择或构造摘要句功能分类任务基准数据集，基于对任务难度和计算成本的综合考量，选择合适的调优方法，形成面向特定功能或特定任务的专用大模型。

在该框架的指导下，本文将大模型驱动的学术文本挖掘研究总结为以下几步：①探究当前主流大模型对于科技文献内容挖掘与知识利用的能力边界，并对比不同的模型应用策略对其学术文本挖掘能力的提升效果，总结适用于科技情报领域的推理端指令策略，从而回答大模型是否有效、怎样有效的问题；②在达到模型推理端的能力上限后，构建或改造面向下游任务的指令数据集，运用各类微调策略，探索其在不同任务上的微调效果，从而回答大模型是否要调优以及如何调优的问题；③在从推理端和调优端两个方面完成的大模型领域适配后，需要探索大模型的科技情报专业能力的形成机制，探索不同学术文本挖掘任务之间的影响与关联，形成科技情报领域下游任务自适应的大模型，从而回答大模型为何有效的问题；④总结面向不同场景和不同任务的最优策略，构建大模型驱动的学术文本挖掘引擎。

本文拟从推理端进行初步探究：从上下文学习、思维链推理等角度，设计推理端指令策略，构建覆盖摘要句功能分类、章节功能识别、引文功能识别、科技实体抽取、科技文本推理和关键词生成6项任务的大模型学术文本挖掘专业能力评测框架，并选取7个国内外主流的指令调优模型，应用该框架评估其在不同维度的专业能力。通过对推理端指令策略和大模型学术文本挖掘专业能力的对比分析，期望能为大模型在科技情报领域的高效应用提供实践参考①。

1 相关工作

1.1 预训练语言模型在科技情报领域的应用

基于高并行Transformer[9]架构的BERT（bidirec‐tional encoder representations from transformers)[10]引发了“预训练-微调”范式的广泛应用，在科技情报领域，许多学者基于这一范式，运用不同的模型结构和微调策略，将预训练语言模型适配各类科技情报领域任务。例如，陆伟等[11]基于预训练语言模型构建了一种面向学术文本的关键词语义功能识别模型，通过将摘要和关键词字段进行向量化表示，对关键词的语义功能进行分类，实现对学术文本中研究问题与研究方法的识别；Jiang等[12]提出了一种可控关键词生成框架，通过将关键词功能添加到输入文本控制模型生成特定类型的关键词，该框架采用预训练-微调范式，对比在Paper with Code数据集上进行微调的Transformer、BART（bidirectional and auto-regressive transformers）、T5（text to text transfer transformer）等生成式模型，验证了该框架在关键词生成任务上的有效性。

然而，传统的面向下游任务的微调方法难以摆脱对标注数据的依赖，无法应对低资源、小样本的任务场景。OpenAI对上下文学习的探索[13]引发了科技情报领域对小样本学习和提示学习的关注。Liu等[14]提出了一种半监督的混合提示学习（mix prompt tuning,MPT）方法，面向计算机领域和生物医学领域的学术文本，在低资源场景下评估了提示学习模型在引文功能、摘要句功能和关键词功能3项任务上的表现，取得了优于微调模型和其他半监督学习模型的结果。大模型在信息推理、数据整编和知识库构建的能力推动了信息资源建设的升级和信息组织模式的转型[15]。张恒等[16]利用ChatGPT(chat generative pre-trained transformer）对SciBERT的训练数据进行增强，显著提升了该模型在研究流程段落识别任务上的分类性能。Wang等[17]提出让大模型通过阅读科技论文产生新的科研想法，设计了一个基于科学假设上下文的“灵感”生成框架，并运用GPT-4对该框架进行评估，探索了一种启发式提示策略，提升了大模型生成内容的技术深度和创新性。

可以发现，在科技情报领域，学者对大模型的应用大多发生在推理端，即运用各种指令策略引导模型解决任务。然而，各类大模型在科技文本上的语义理解能力有何差异，如何从推理端发挥大模型在科技情报领域的最大效能，当前尚未有研究对此进行实验论证和结果探讨。对大模型学术文本挖掘能力边界和推理端指令策略的探索，是当前亟待解决的研究问题。

1.2 大模型推理端应用策略及能力评测

大模型推理端的应用主要涉及上下文学习、思维链推理等指令工程技术[18]。上下文学习最早由OpenAI在GPT-3论文中提出，即通过将任务描述、任务样例拼接为格式化的自然语言指令作为模型的输入文本[13]，令模型通过文本续写的方式完成指令要求。上下文学习的本质是充分掌握蕴含在模型参数内的技巧和任务，因此，随着模型规模的扩大，模型的上下文学习能力会逐渐增强。此外，指令微调也能够提升模型的上下文学习能力。相比于上下文学习通过以“输入-输出对”构建样例进行提示，思维链推理是通过在指令与答案之间构建一系列推理步骤来提示模型生成最终输出，从而提升语言模型在推理任务上的表现。Kojima等[19]设计了一种零样本思维链策略，通过在模型生成答案之前添加“Let’s think step by step”，即可提升模型在各类推理任务上的性能表现。Wei等[20]发现，思维链推理仅对参数规模超过100亿的大模型有作用，无法提升小模型的推理性能，证实了思维链推理是语言模型超过一定规模时涌现的特殊能力。这为本文采用合适的指令策略引导模型解决科技情报领域问题提供了借鉴。

当前，对大模型能力的评测包含通用能力评测和专业能力评测。Qin等[21]搜集了20个通用领域数据集，涵盖7类自然语言处理任务，对ChatGPT的通用能力进行评估，发现大模型在推理类和对话类的通用任务上表现良好，但在命名实体识别等序列标注任务上仍然面临挑战，且在大多数情况下，ChatGPT的表现不如特定任务的微调模型。Nori等[22]对提示工程进行了系统探索，设计了面向医学领域的提示策略Medprompt，该策略提升了GPT-4的专业能力，并在9个医学领域基准数据集上取得了SOTA

(state-of-the-art）效果，且优于微调模型，如Med-PALM 2 (medical pretrained language model2）的效果。张颖怡等[23]运用提示学习方法，从实体识别、伪标签生成和训练数据生成3个视角出发，评估了ChatGPT在学术论文实体识别任务上的可用性，对大模型在科技情报领域的专业能力进行了初步评测。

整体而言，当前科技情报领域对大模型的应用仍处于初级阶段，尚未形成规范化的指令策略构建方法，大模型学术文本挖掘专业能力评测框架不够完善，对不同参数规模通用大模型对学术文本的语义理解能力缺乏整体认知。针对上述研究不足，本文设计了面向学术文本挖掘任务的推理端指令策略，构建了涵盖关键词生成、科技实体抽取、科技文本推理、章节功能识别等科技情报典型任务的大模型学术文本挖掘专业能力评测框架，对当前主流的国内外指令调优模型在推理端的学术文本挖掘专业能力进行评测，为大模型与科技情报领域的深层嵌入提供实践参考。

2 推理端指令策略构建及模型专业能力评测

大模型能否仅通过指令引导便成为“科技情报领域专家”？何种指令策略能够更大程度地发挥大模型在学术文本挖掘任务上的性能？为了回答这些问题，本文在不调整模型参数的条件下，从上下文学习、思维链推理等角度运用不同的指令策略，对当前国内外主流的大模型在学术文本挖掘相关任务上的专业能力进行评测，并依据各个模型在不同指令策略下的指标表现，总结不同模型适用于学术文本挖掘的指令策略，评价模型在不同维度的专业能力。

2.1 大模型学术文本挖掘指令策略

大模型的参数中蕴含着解决特定领域问题的专业能力，这些能力可以通过精确的自然语言指令进行提示和激活[22]。本文对指令工程相关技术进行系统梳理，总结面向学术文本挖掘相关任务的指令策略，按照选择指令模板、编写任务描述、选择样本示例和构造思维链4个步骤，形成面向学术文本挖掘各项任务的自然语言指令，用于从模型参数中挖掘解决科技情报领域问题的专业能力，具体流程如图2所示。

(1）指令模板选择指令模板是指用于填充指令内容、参考文本、任务描述、样本示例各类指令信息的骨架，能够提示大模型生成回应指令内容的输出，从而实现文本续写向多轮对话的转换。不同的基础模型在指令调优时选用的指令模板有所差异，但其作用都是将用户指令嵌入模板中，从而实现对指令文本向任务答案之间映射的拟合。对于不同模型，选择更为适配的指令模板不仅能使其更好地遵循指令，还能提升其在专业任务上的性能表现。例如，对于Alpaca系列模型，本文选用的指令模板为“###Instruction:\n{instruction}\n\n###Input:\n{input}\n\n###Response:”，当将其应用于ChatGLM系列模型时，会导致其无法遵循指令，输出无意义的内容；当将其应用于ChatGPT时，会干扰模型生成格式，降低专业性能表现。由此可见，选择适当的指令模板对于大模型解决科技情报领域专业任务是十分重要的。本文以模型的指令调优模板或官方推理模板为基准，在不同任务上进行小批量测试和调整，以确保指令能够引导模型生成满足任务要求且便于自动评测的结构化或半结构化文本。

(2）任务描述编写

任务描述通常是对任务类型的说明文本，对于不同的问题，任务描述中侧重的内容略有差异。对于学术文本挖掘中的分类问题，任务描述文本中应该包含任务解决方法、输出类型解释、输出内容限定等信息。例如，“摘要句功能分类任务是通过理解论文摘要中的每个句子的功能，将其所属功能类别划分为方法、背景、结果、结论、目的，以下是各个功能的解释……仅输出类别文本”；而对于学术文本挖掘中的抽取问题，任务描述文本中应当在上述内容的基础上着重说明输出格式限制。例如，“请以JSON列表的格式进行输出”。

(3）样本示例选择样本示例在上下文学习中扮演着至关重要的角色，所选的样本应当具有广泛的代表性，从而充分发挥模型在推理端的专业能力，本文在样例选择的过程中遵循以下原则：

第一，若所选测试数据包含训练集或验证集，则本文选择能够覆盖所有标签的数据样例，例如，对于信息抽取类数据集，本文选择尽可能包含所有类型实体的句子作为样本实例，从而引导模型熟悉文本输入与输出标签的对应关系；

第二，若测试数据不存在训练集或验证集，则由多位博士研究生共同编写样本示例，并在领域专家的监督下，确保样例的代表性和与测试数据的相关性。

在任务描述的基础上，若不添加样本示例，则构成零样本（0-shot）指令策略，不同样本数量构成了单样本（1-shot）和少样本（few-shot）两种指令策略。

(4）思维链构建对于文本推理类任务，在模型输入上下文中构建思维链能够激发大模型进行逐步推理，从而提升模型在复杂推理任务上的表现。

在处理科技情报文本相关的推理任务时，本文参考Kojima等[19]和Suzgun等[24]总结的思维链方法，设计了面向科技情报领域的零样本思维链（zeroshot chain-of-thought,0-shot CoT）和少样本思维链（few-shot chain of thought,few-shot CoT）。具体来说，对于学术文本挖掘任务，零样本思维链将任务描述中的输出约束改为“请你一步一步思考，并逐步输出推理过程”。当零样本思维链引导各类模型生成推理结果后，人工对输出内容中的推理过程和推理结果进行校验，选择正确且具有代表性的示范样本作为少样本思维链的示例，基于此进行测试，并对比不同的思维链策略对大模型科技情报文本推理能力的提升效果。

2.2 学术文本挖掘专业能力评测框架

为了细致刻画当前主流大模型的学术文本挖掘能力边界，本文构建了一个面向科技情报领域的评测框架，如图3所示。该框架按照“能力-任务-指标”划分了3层维度，共包含6个评测数据集，总计37747条测试样本。

2.2.1 能力维度

为了更加全面和准确地评测大模型的学术文本挖掘能力，本文参照文本挖掘技术的任务划分方式，从知识单元粒度和能力本质特征角度，将各类学术文本挖掘任务归纳为文本分类、信息抽取、文本推理和文本生成4个类别，每个类别的任务都具有各自的特点和要求，能够用于评估大模型在不同维度的学术文本挖掘专业能力。

4项学术文本挖掘专业能力在知识粒度和任务要求等方面各有不同：文本分类能力侧重于大模型对于学术文本中句子层级和段落层级知识单元的理解，如摘要、章节标题、章节段落等，通常是对一种或多种类型的知识单元进行属性或关系的判别；信息抽取能力重点关注词汇层级的知识单元，包括学术文本中的单词、关键词、命名实体、概念和主题等，旨在表示大模型从非结构化或半结构化的数据中提取有用的结构化信息的水平，该能力可用于构建复杂的科技情报知识图谱，或用于更高级别的数据分析和推理；文本推理能力主要面向句子层级的知识单元，是指大模型理解学术文本关键句之间隐含逻辑关系或事实的能力，在自动问答系统、对话理解和推荐系统等应用场景中尤为重要；文本生成能力则更多地关注如何从现有的知识单元中派生或

重组出新的知识单元。

为了对大模型的学术文本挖掘能力进行评测，需要选取各项关键能力的代表性任务，开展大模型在科技情报领域典型任务上的性能分析和指标评测，从而回答大模型能否成为“科技情报领域专家”的问题。

2.2.2 任务维度

科技情报领域典型任务的识别是本文探究大模型在该领域应用效能与前景的基础。对于不同学术文本挖掘能力的评测，由于研究对象的差异，需要选择适当的任务来衡量大模型在不同能力维度的性能表现。例如，衡量大模型文本分类能力的任务通常包括情感极性分析、论文章节功能识别、引文功能识别等，信息抽取能力通常涉及关键词抽取、科技实体抽取和科技事件抽取等任务，文本推断能力通常包括因果关系推断、文本蕴涵和观点挖掘等任务，文本生成能力包括但不限于关键词生成、学术文本摘要生成、一句话摘要生成和风格化摘要生成等任务。

大模型应用于学术文本挖掘任务的出发点是发挥其通用的任务解决范式和专业文本理解能力，对学术文本全粒度知识进行加工，从而更好地为学者或用户服务。因此，本文遵循以下两个原则选取评测数据：一是覆盖学术文本知识单元全粒度，包括关键词、句子、摘要、段落、章节、引文等；二是有明确可信、能够准确反映人类偏好的评测指标可供使用。由此，本文选取了学术文本挖掘相关数据集用于模型评测，英语是目前主流的大模型均支持的语言。为了保证大模型科技情报专业能力评测的公平性，本文均选用英文数据集作为评测基准，数据集基本情况如表1所示。

本文使用各个数据集的全量测试集，将其转化为适用于不同模型的自然语言指令，从而保证评测结果的可靠性。在科技情报领域，典型的文本分类任务包括摘要句功能分类、章节功能识别、引文功能识别等，本文分别选择RCT-20K、ScienceDirect和ACL-ARC作为基准数据集，从多个维度评估大模型在科技情报领域的文本分类能力。在信息抽取方面，对SciERC的数据进行改造，作为评估科技实体抽取任务的基准数据集。对于文本推理类任务，本文选择SciNLI作为评估模型在科技情报领域推理能力的基准数据集，SciNLI是一个面向学术文本的自然语言推理数据集，共包含自然语言处理与计算语言学领域科学论文的10.7万条句子对。对于文本生成类任务，本文以科技情报领域的“关键词生成”为典型任务对大模型的进行评测，沿用Jiang等[12]提出的可控关键词生成框架与PwC数据集，基于论文标题和摘要生成特定类型的关键词。这些任务通常具有统一的评价体系和评测指标，便于量化和对比不同模型运用不同指令策略的性能表现，从而验证大模型在科技情报领域的能力边界。

2.2.3 指标维度

为了消除人类主观偏见对评测结果产生的误差，同时确保低资源设定下指令引导的大模型与“预训练-微调”模型的可比较性，本文选择各类任务的典型指标用于评测。例如，对于文本推理任务，使用准确率（accuracy,Acc）与宏平均F1值（marco F1）作为评测指标。由于大模型的输出结果难以控制，对于指令遵循能力较弱的小参数模型（如Al - paca-7B）尤为明显。因此，本文在设计指令引导模型生成结构化文本的基础上，对于未遵循指令生成非结构化文本的模型输出，运用规则匹配等方法对输出文本进行后处理，以实现对生成文本的自动评测。

特别地，若作者在提出基准数据集时有额外的指标限定（如取模型输出的前5个样本作为预测值计算精确率（P@5）和召回率（R@5），或计算F1时使用加权宏平均等），则为了保证与基线对比的公平性，本文以作者提出的指标为准。

2.3 学术文本挖掘专业能力评测对象

大模型可以分为基础模型和指令微调模型。指令微调能够提升大模型的上下文学习和思维链推理能力。为了在推理端最大程度地挖掘大模型的学术文本挖掘能力，同时考虑到目前缺失科技情报领域特有的大模型，本文从来源国家和开源情况两个维度，选用当前主流的指令微调模型作为评测对象，分别选择Alpaca[30]、ChatGPT[31]、ChatGLM[32]、文心一言（ERNIE Bot)[33]等4个模型系列分别作为国外开源模型、国外私域模型、国内开源模型、国内私域模型的代表。所选指令调优模型及具体参数情况如表2所示。

3 实验与分析

本文所有涉及大模型专业能力评测的实验均在Ubuntu 20.04操作系统中完成，使用4张NVIDIA Tesla A100-80G，在Pytorch 2.0.0框架下进行模型推理测试。

对于7种主流的指令调优模型，本文分别测试其在不同指令策略下，4类共6项学术文本挖掘能力的性能表现，共包含140组推理端实验，本文按照4个能力维度对实验结果进行分析。

3.1 文本分类能力评测

如表3所示，对于RCT-20K和ScienceDirect数据集，均采用加权宏平均F1值（weighted macro F1）进行评测，ACL-ARC数据集则采用宏平均F1值作为评测指标。GPT-4是分类任务上表现最优的模型，在摘要功能分类、章节功能识别和引文功能识别3类任务上的最佳调和平均值分别可以达到73.06%、54.30%和61.06%。通过对比不同指令策略在3个分类数据集上的性能，我们认为，少样本学习策略在分类任务上是行之有效的，不同大模型运用零样本（0-shot）、单样本（1-shot）和少样本（few-shot）3种指令策略在3类任务上的平均调和平均值分别为35.99%、36.51%和40.35%。

对于Alpaca-7B和ChatGLM-6B这类参数规模较小的模型，单样本和少样本带来的指标提升效果十分不稳定。例如，在摘要句功能分类任务上，Al - paca-7B在单样本设定下的F1为3.35%，远低于零样本的32.38%。为了探究其原因，本文对摘要句功能的推理结果进行了统计，发现单样本设定下的大多数模型分类结果与样本示例的分类标签一致。例如，当样本示例的分类标签为

“CONCLUSION”时，在30135个测试样本中，Alpaca-7B模型将27622个样本的章节功能预测为“CONCLUSION”，而其中预测正确的样本仅为2301个。由此可知，单样本指令策略对任务理解能力不足的Alpaca-7B模型效果较差，在进行指令设计时，应尽量避免使用单样本来引导此类模型解决科技情报领域的文本分类问题。

3.2 信息抽取能力评测

如表4所示，对于信息抽取任务，本文选用微平均F1值（micro F1）作为评测指标。从横向来看，随着样本数量的增多，各个模型的科技实体识别能力呈现增强的趋势。零样本、单样本和少样本3类策略在所有模型上的平均F1值分别为10.03%、13.06%和17.44%，这表明样本示例对于模型解决信息抽取类任务具有显著的指导作用。从纵向来看，规模更大的模型在科技文本识别任务上表现更好，GPT-4在科技文本通用实体抽取任务上表现最好，在零样本、单样本和少样本场景下的F1值分别为21.30%、22.76%和24.80%。对于部分模型，单样本设定下的抽取表现相比于零样本设定没有显著提升，甚至弱于零样本效果，且这一现象在千亿参数规模的大模型上更为常见。例如，GPT-3.5-turbo与ERNIE Bot在科技文本实体抽取任务中零样本设定的效果均优于单样本效果。

随着模型参数规模的增大，样本增多所带来的实体抽取效果提升逐渐减小。各个模型的F1指标样本提升率如图4所示。可以看出，对于千亿级参数规模的大模型，GPT-4的提升率最高，为16.43%；而对于百亿级参数规模的模型，提升率均在100%以上，远高于千亿级模型的提升率。本文对这一现象进行了深入分析，其原因可归纳为两个方面：一方面，零样本设定下千亿级参数规模模型的抽取指标更高，指标提升的难度更高，导致增加样本带来的指标提升率逐渐减小；另一方面，超大规模模型的训练语料更丰富，选用的数据样例带来的增量知识与在模型本身蕴含的存量知识重叠概率更高，因此，提升率较百亿级模型显著降低。

3.3 文本推理能力评测

如表5所示，对于文本推理任务，本文选用准确率和宏平均F1值作为评测指标，基于零样本、单样本、少样本、零样本思维链（0-shot CoT）和少样本思维链（few-shot CoT）5种指令策略对模型的学术文本推理能力进行评测。从横向对比来看，虽然5种指令策略在不同模型上各有优劣，但总体来看，思维链策略在推理类任务上存在如下优势：零样本思维链相比于零样本指令策略在所有模型上的平均准确率由36.85%提升至36.92%，少样本思维链相比于少样本策略的平均准确率由36.66%提升至37.92%。然而，样本示例在该任务上的效果并不显著，单样本、少样本策略相较于零样本在准确率指标上均未有提升。从纵向对比来看，GPT-4在该任务上仍然是最优模型，平均准确率为55.44%;ER‐NIE Bot次之，ChatGLM2位列第三，且超过了GPT-3.5-turbo，平均准确率达到了38.93%。

具体来看，零样本思维链策略对于Alpaca、ChatGLM两类十亿级别的指令调优模型的学术文本推理能力提升效果较弱，而对千亿级别的调优模型的提升效果较明显。其中，ERNIE Bot的提升幅度最大，准确率由40.90%提升到了46.73%，提升率为14.25%；与之相对应的是，少样本思维链对于Alpaca-13B、ChatGLM等百亿级别参数模型的提升效果较明显，提升幅度最大的模型为Alpaca-13B，相较于零样本策略，准确率的提升率为10.30%。

3.4 文本生成能力评测

如表6所示，对于关键词生成任务，本文选用F1@5作为评测指标。当前主流的大模型在基于论文题目与摘要生成不同类型关键词这一任务场景下的表现并不理想，即使效果最好的模型ERNIE Bot，其F1@5也仅在少样本设定下达到11.90%。7.7亿参数规模的T5-Large模型在经过微调后，F1@5能够取得55.8%的性能，远高于万亿级参数规模的大模型在少样本场景下取得的效果[12]。

PwC-kw数据集中论文的关键词由Paper With Code的标签转化而来，因而所有关键词都是预定义且有限的，T5等“小模型”在训练集上进行微调的过程中，能够学习到关键词的表述方式和定义规则，这些知识在自然语言指令中是难以显式表达的。例如，在表7中，粗体代表与标签完全一致的关键词；斜体代表与标签含义相同，但表述略有差异，即使进行词根还原也与标签无法匹配的关键词。该样例中的指标类关键词为“Accuracy (CS)”和“Accuracy (CV)”，其并未出现在标题和摘要中，而是在训练集中出现多次，因而指令引导下的ER‐NIE Bot的仅能预测为“Accuracy”，在训练集上微调后的模型能够学习到该关键词的准确表述，并输出准确结果。

对于分类、抽取、推理等输出结果具有直接约束的任务，该现象表现得并不明显，而对于难以通过指令覆盖所有规则的生成类任务，其数据评测结果往往并不尽如人意。因此，本文认为当前并非所有的学术文本挖掘任务都可以仅在推理端依靠大模型解决，由于通过学习大量数据样例才能学习到的“隐知识”难以通过指令描述，模型调优仍然是不可替代的。

4 讨论

4.1 指令策略对比分析

通过第3节对主流大模型在6项具体任务上的评测实验，本文对大模型在不同指令策略下的性能表现进行了对比，总结不同模型在各类学术文本挖掘任务上的最优指令策略，如图5所示。

在摘要句功能分类、章节功能识别、引文功能识别等分类任务上，Alpaca-7B、Alpaca-13B、Chat‐GLM存在不需要任何样本示例就取得各自最优性能的情况，即样本示例对这类模型分类性能的提升效果不够稳定；而对于ChatGLM2、GPT-3.5-turbo、ERNIE Bot等模型，大模型通常在样本数量更多的情况下取得最优结果。正如Brown等[13]在“Lan‐guage Models Are Few-Shot Learners”一文中的推测：上下文学习涉及吸收模型参数内的各项技能，参数规模越大的模型表现出的上下文学习能力越强。该现象在科技情报领域中同样适用，对于输出文本搜索空间较小的分类任务，参数较小的模型往往难以从示例样本中学习到任务解决方法，训练策略更优（如ChatGLM）和参数规模更大（如GPT-3.5-tur‐bo等）的模型则具备更强大的上下文学习能力。

在科技实体抽取和关键词生成任务上，所有模型均在存在样本示例的情况下取得了各自最优的结果。本文对Alpaca-7B、ChatGLM等模型的输出内容进行采样，发现这类模型的输出结果存在格式混乱的问题，导致内容无法解析，使得指标评测失效。由于抽取任务的输出文本格式相较于分类任务更为复杂，对于参数较少的模型来说，仅通过任务描述难以引导模型生成符合格式要求的输出文本。在这种情况下，少量示例更能引导模型生成正确格式的内容，从而提升了抽取任务的指标性能。

科技文本推理任务要求模型具备理解和分析科技工作的能力，涉及对抽象知识和复杂概念的逻辑关系推理。对于GPT-4、ERNIE Bot等参数量更大的模型，连贯的思维链条能够增强其对复杂科技文本逻辑框架构建和推理过程，从而更准确地推断出正确结果；对于Alpaca-7B和ChatGLM2，思维链策略的提升效果并不显著，有时反而会降低模型的推理性能。

综上所述，不同模型在各类学术文本挖掘任务上表现出了不同的最优指令策略，反映了不同模型处理复杂任务时的适应性和逻辑性，指令策略选择应当根据实际任务需求、任务推理难度和可解释要求等来确定。例如，十亿级参数规模模型在分类任务上对样本示例的提示并不敏感，但在抽取、生成等较为复杂的任务上样本示例效果显著增强。随着参数规模增大，模型逐渐展现出了对上下文学习的依赖，各类千亿级大模型均在单样本或少样本设定下表现出了最优效果，且在思维链策略下表现出了更强大的推理效果。

4.2 模型能力对比分析

4.2.1 不同参数规模大模型能力对比

为了更直观地展现大模型在不同样本设定下不同能力维度的效果，本文将各个数据集上模型的核心指标进行归一化，映射到同一区间内，构建大模型的6项专业能力值，计算方法为其中，CapabilityD(KM)是模型M在数据集D上表现出的基本能力；对于数据集D,KM为评测模型M的核心指标（F1值或准确率）；Kmin为各个模型指标的极小值；Kmax为各个模型指标的极大值；α和β分别为归一化区间的下界和上界，为了便于展示，本文取α=0.1,β=1。基于公式(1)，本文计算了各个模型的6项学术文本挖掘能力值，由于单样本选择对模型的扰动较大，本文仅对零样本和少样本场景的模型能力值进行可视化，如图6所示。

模型参数规模带来的能力涌现在科技情报领域同样适用，千亿级规模模型的各项能力全面领先于百亿级模型。对于Alpaca-7B、ChatGLM2-6B等参数规模较小的模型，少样本策略下的模型能力值整体高于零样本策略，这意味着通过在上下文中增加示范样例能够增强十亿及百亿级模型的学术文本挖掘能力，并缩小了其与千亿模型之间的差距。但整体来看，推理端的各类策略对规模较小的模型的效果仍然收效甚微，若要充分发挥百亿以下参数规模大模型的潜在能力，仍需借助外部知识库扩展模型输入，或通过调整模型参数，以适应科技情报领域的专业问题。值得一提的是，在少样本设定下的关键词生成任务与零样本设定下的引文功能识别任务上，百度开发的ERNIE Bot与清华智谱的ChatGLM在指标上分别超过了业界公认最优的GPT-4模型。

4.2.2 大模型与传统深度学习模型能力对比

如表8所示，首先，在所有任务上，均由GPT-4、ERNIE Bot等规模更大的模型取得了推理端的最优结果，这意味着模型参数增多带来的能力涌现现象同样发生在学术文本挖掘领域，更大的参数量使得模型能够更好地依照指令对非结构化的学术文本进行深度理解和语义组织。其次，在摘要句功能分类、章节功能识别、科技实体抽取、科技文本推理和关键词生成5类任务上，模型均在更复杂的指令策略下取得了最优结果。Nori等[22]研究发现，多种指令策略的复杂组合能够释放千亿级参数规模大模型更深层次的专业能力，在某些任务上能够表现出优于全样本微调的效果。

然而，相比于“预训练-微调”范式下的“小模型”，基于指令工程的大规模语言模型在学术文本挖掘各项任务的表现仍然较弱，即推理端的指令引导难以充分发挥大模型在科技情报领域的效能，仅依靠自然语言指令的知识调度不足以实现大模型与科技情报领域的深度融合，仍需在调优端对模型进行参数调优，以实现应对不同领域的专业化适配。

5 总结与展望

本文以学术文本挖掘任务为切入点，设计了少样本、思维链等推理端指令策略，构建了涵盖4个能力维度共6项任务的大模型学术文本挖掘专业能力评测框架，并选取了7个国内外主流的指令调优模型进行实验，对比了不同参数模型应用零样本、少样本以及思维链等指令策略在不同任务上的指标表现。实验结果表明，千亿级参数规模的大模型，在经过适当的指令引导后能够表现出可观的效果，但仍弱于“预训练-微调”范式下的传统模型，即面向科技情报领域，推理端的指令策略效果有限，仍需在调优端对模型参数进行领域化适配。

未来，可继续深入探索大模型在科技情报领域的应用路径：基于各类指令调优策略与人类意图对齐策略对当前基础模型进行优化，探索大模型学术文本挖掘能力的形成机制，形成科技情报领域下游任务自适应的大模型。基于此，进一步研发大模型驱动的科技情报细粒度解析平台，构建面向信息资源管理领域的服务引擎，提供端到端的科技文献智能服务。

【参考文献】

[1]张智雄,于改红,刘熠,等. ChatGPT对文献情报工作的影响[J].数据分析与知识发现, 2023, 7(3):36-42.

[2] Huang S Z, Qian J J, Huang Y, et al. Disclosing the relationship between citation structure and future impact of a publication[J].Journal of the Association for Information Science and Technology, 2022, 73(7):1025-1042.

[3]王鑫,程齐凯,马永强,等.基于层次注意力网络的论证区间识别研究[J].情报工程, 2020, 6(3):52-62.

[4]程齐凯,李信,陆伟.领域无关学术文献词汇功能标准化数据集构建及分析[J].情报科学, 2019, 37(7):41-47.

[5] Ma Y Q, Liu J W, Lu W, et al. From “what” to “how”:extracting the procedural scientific information toward the metric-optimization in AI[J]. Information Processing&Management, 2023, 60(3):103315.

[6]陆伟,马永强,刘家伟,等.数智赋能的科研创新——基于数智技术的创新辅助框架探析[J].情报学报, 2023, 42(9):1009-1017.

[7]陆伟,汪磊,程齐凯,等.数智赋能信息资源管理新路径:指令工程的概念、内涵和发展[J].图书情报知识, 2024, 41(1):6-11.

[8] Ding N, Qin Y J, Yang G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature Machine Intelligence, 2023, 5(3):220-235.

[9] 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. Red Hook:Curran Associates,2017:6000-6010.

[10] Devlin J, Chang M W, Lee K, et al. BERT:pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. Stroudsburg:Association for Computational Linguistics, 2019:4171-4186.

[11]陆伟,李鹏程,张国标,等.学术文本词汇功能识别——基于BERT向量化表示的关键词自动分类研究[J].情报学报, 2020,39(12):1320-1329.

[12] Jiang Y, Meng R, Huang Y, et al. Generating keyphrases for readers:a controllable keyphrase generation framework[J]. Journal of the Association for Information Science and Technology, 2023, 74(7):759-774.

[13] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems,2020, 33:1877-1901.
[14] Liu J W, Xiong Z, Jiang Y, et al. Low-resource multi-granularity academic function recognition based on multiple prompt knowledge[OL].(2023-05-05)[2023-12-25]. https://arxiv.org/pdf/2305.03287.
[15]陆伟,刘家伟,马永强,等. ChatGPT为代表的大模型对信息资源管理的影响[J].图书情报知识, 2023, 40(2):6-9, 70.
[16]张恒,赵毅,章成志.基于SciBERT与ChatGPT数据增强的研究流程段落识别[J].情报理论与实践, 2024, 47(1):164-172, 153.
[17] Wang Q Y, Downey D, Ji H, et al. Learning to generate novel scientific directions with contextualized literature-based discovery[OL].(2023-10-12)[2023-12-25]. https://arxiv.org/pdf/2305.14259v3.
[18] Zhao W X, Zhou K, Li J Y, et al. A survey of large language models[OL].(2023-11-24)[2023-12-25]. https://arxiv.org/pdf/2303.18223.
[19] Kojima T, Gu S S, Reid M, et al. Large language models are zeroshot reasoners[J]. Advances in Neural Information Processing Systems, 2022, 35:22199-22213.
[20] Wei J, Wang X Z, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the 36th Conference on Neural Information Processing Systems. Cambridge:MIT Press, 2022:24824-24837.
[21] Qin C W, Zhang A, Zhang Z S, et al. Is ChatGPT a general-purpose natural language processing task solver?[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg:Association for Computational Linguistics, 2023:1339-1384.
[22] Nori H, Lee Y T, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine[OL].(2023-11-28)[2023-12-25]. https://arxiv.org/pdf/2311.16452.
[23]张颖怡,章成志,周毅,等.基于ChatGPT的多视角学术论文实体识别:性能测评与可用性研究[J].数据分析与知识发现,2023, 7(9):12-24.
[24] Suzgun M, Scales N, Schärli N, et al. Challenging BIG-bench tasks and whether chain-of-thought can solve them[C]//Proceedings of the 61st Annual Conference of the Association for Computational Linguistics. Stroudsburg:Association for Computational Linguistics, 2023:13003-13051.
[25] Dernoncourt F, Lee J Y. PubMed 200k RCT:a dataset for sequential sentence classification in medical abstracts[C]//Proceedings of the 8th International Joint Conference on Natural Language Processing. Stroudsburg:Association for Computational Linguistics, 2017:308-313.
[26]王佳敏,陆伟,刘家伟,等.多层次融合的学术文本结构功能识别研究[J].图书情报工作, 2019, 63(13):95-104.
[27] Bird S, Dale R, Dorr B J, et al. The ACL anthology reference corpus:a reference dataset for bibliographic research in computational linguistics[C]//Proceedings of the Sixth International Conference on Language Resources and Evaluation. Stroudsburg:Association for Computational Linguistics, 2008:1755-1759.
[28] Luan Y, He L H, Ostendorf M, et al. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg:Association for Computational Linguistics, 2018:3219-3232.
[29] Sadat M, Caragea C. SciNLI:a corpus for natural language inference on scientific text[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg:Association for Computational Linguistics, 2022:7399-7409.
[30] Taori R, Gulrajani I, Zhang T Y, et al. Alpaca:a strong, replicable instruction-following model[EB/OL].[2023-07-18]. https://crfm.stanford.edu/2023/03/13/alpaca.html.
[31] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C]//Proceedings of the36th International Conference on Neural Information Processing Systems. Red Hook:Curran Associates, 2022:27730-27744.
[32] Du Z X, Qian Y J, Liu X, et al. GLM:general language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg:Association for Computational Linguistics, 2022:320-335.
[33] Sun Y, Wang S H, Feng S K, et al. ERNIE 3.0:large-scale knowledge enhanced pre-training for language understanding and generation[OL].(2021-07-05)[2023-07-18]. https://arxiv.org/pdf/2107.02137.
[34] Cohan A, Ammar W, van Zuylen M, et al. Structural scaffolds for citation intent classification in scientific publications[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. Stroudsburg:Association for Computational Linguistics, 2019:3586-3596.

5. 复制比按照“四舍五入”规则,保留1位小数;若您的文献经查重检测，复制比结果为0，表示未发现重复内容，或可能存在的个别重复内容较少不足以作为判断依据
6. 红色文字表示文字复制部分;绿色文字表示引用部分(包括系统自动识别为引用的部分);棕灰色文字表示系统依据作者姓名识别的本人其他文献部分
7. 系统依据您选择的检测类型(或检测方式)、比对截止日期(或发表日期)等生成本报告
8. 知网个人查重唯一官方网站:https://cx.cnki.net