doi:10.3772/j.issn.2095-915x.2024.05.009

开放科学（资源服务）标识码（OSID）

# 基于大模型的科研设备成本评估框架

程齐凯[1,2] 雷道宇[1,2] 石湘[1,2] 刘寅鹏[1,2]

1. 武汉大学信息管理学院 武汉 430072；
2. 武汉大学信息检索与知识挖掘研究所 武汉 430072

**摘要：**[ 目的 / 意义 ] 提出一种创新的基于大模型的科研设备成本评估框架，旨在解决传统成本评估方法中的局限性，如成本评估的不精确性和效率低下问题。自动化地从科研论文中抽取实验材料与设备信息，并设计科研设备成本估算模型，从而精准和高效地评估科学研究成本，为实验成本的精确评估和科研资源的有效利用提供了新的工具和方法。[ 方法 / 过程 ] 以物理和计算机领域为例，利用 arXiv 数据库与 Paper With Code 网站提供的论文数据构建了一个训练数据集，并采用 LoRA 微调技术在基准模型 LLaMA2-13b 上进行微调，使其能够精确抽取目标领域论文中关于实验设备与材料的详细信息。通过 Wikipedia 进行实体链接消歧，并综合考虑材料设备的价格波动，设计了一种平均情况分析的成本估算公式，以计算机视觉领域为例对科研设备成本评估框架的有效性进行验证。[ 局限 ] 只在计算机领域和物理领域进行了实验，同时数据集的构建主要依赖于公开可获取的论文数据，这可能限制了成本评估框架的泛化能力和准确性。[ 结果 / 结论 ] 通过对计算机科学与物理学领域的科研论文进行实证分析，展示了基于大模型的科研设备成本评估框架的有效性。通过 LoRA 技术微调的 LLaMA2 模型在信息抽取任务上显示出较高的准确率和召回率，证明了本框架在精准抽取实验材料与设备信息方面的能力。同时，在计算机视觉领域开展了成本估算分析，揭示了计算资源已经成为制约计算机视觉领域科研产出的关键因素之一和特定的算法模型结构或研究范式存在性能上限等结论。这些发现与实际科研活动相吻合，证明了本文提出的成本评估框架能够准确反映科研实践的现实情况，为科研项目的资源优化提供了重要参考。

**关键词：**信息抽取；大模型；高效微调；成本估算框架

**中图分类号：**G35；TP391

# Scientific Research Equipment Cost Evaluation Framework Based On Large Language Model

CHENG Qikai[1,2] LEI Daoyu[1,2] SHI Xiang[1,2] LIU Yinpeng[1,2]

**基金项目** 国家自然科学基金面上项目“基于机器阅读理解的科学命题文本论证逻辑识别”（72174157）；国家自然科学基金重点项目“数智赋能的科技信息资源与知识管理理论变革”（72234005）。

**作者简介** 程齐凯（1989-），博士，副教授，主要研究方向为文本挖掘、信息检索；雷道宇（1998-），硕士研究生，主要研究方向为自然语言处理，E-mail: leidaoyu@whu.edu.cn；石湘（1998-），博士研究生，主要研究方向为信息检索、知识管理；刘寅鹏（1998-），博士研究生，主要研究方向为文本挖掘、自然语言处理。

**引用格式** 程齐凯 , 雷道宇 , 石湘 , 等 . 基于大模型的科研设备成本评估框架 [J]. 情报工程 , 2024, 10(5): 99-114.

doi:10.3772/j.issn.2095-915x.2024.05.009

1. School of Information Management, Wuhan University, Wuhan 430072, China;
2. Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, China

**Abstract:** [Objective/Significance] This study proposes an innovative framework for assessing the cost of scientific research equipment based on large language models, aiming to address the limitations of traditional cost assessment methods, such as the inaccuracy and inefficiency of cost estimation. By automatically extracting experimental material and equipment information from scientific research papers and designing a cost estimation model for research equipment, this framework provides a new tool and method for accurately and efficiently evaluating the cost of scientific research, enabling precise cost assessment and effective utilization of research resources. [Methods/Processes] Using physics and computer science as examples, this study constructs a training dataset based on the paper data provided by the arXiv database and the Paper with Code website. It employs the LoRA fine-tuning technique on the benchmark model LLaMA2-13b, enabling it to accurately extract detailed information about experimental equipment and materials from papers in the target domains. Entity linking disambiguation is performed using Wikipedia, and a cost estimation formula based on average-case analysis is designed, considering the price fluctuations of materials and equipment. The effectiveness of the research equipment cost assessment framework is validated using the field of computer vision as an example. [Limitations] Experiments were conducted only in the computer science and physics domains, and the construction of the dataset primarily relies on publicly available paper data, which may limit the generalizability and accuracy of the cost assessment framework. [Results/Conclusions] Through empirical analysis of scientific research papers in the fields of computer science and physics, this study demonstrates the effectiveness of the research equipment cost assessment framework based on large language models. The LLaMA2 model fine-tuned using LoRA technology exhibits high accuracy and recall in the information extraction task, proving the framework's ability to accurately extract experimental material and equipment information. Additionally, the study conducts cost estimation analysis in the field of computer vision, revealing that computational resources have become one of the key factors constraining research output in computer vision, and that specific algorithmic model structures or research paradigms have performance limits. These findings align with real-world scientific research activities, demonstrating that the proposed cost assessment framework can accurately reflect the realities of scientific practice and provide important references for optimizing resources in research projects.

**Keywords:** Information Extraction; Large Language Model; Efficient Fine-tuning; Cost Evaluation Framework

## 引言

科学实验是科研工作者探索未知、验证假设、积累知识的核心手段，也是推动科学理论发展和技术创新的直接动力。科技文献承载了研究者的思想和成果[1]，实验则是这些成果诞生和检验的实践场域。每项科学实验都伴随着一系列成本的评估，包括实验材料、设备和时间等资源的投入，这些成本不仅会影响研究的可行性和实验的规模，也是科研项目规划和决策过程中不可或缺的考量因素。尤其在一些高度专业化和技术密集的领域[2]，如物理和计算机领域，实验设备的成本和配置问题显得更为复杂，在进行研究之前必须采用更加精细化和动态化的成本评估方法，以确保科研资源的有效利用和项目的顺利进行。

近年来，信息技术的发展促进了成本管理工具的创新，一些研究通过开发软件和算法来帮助科研人员估算项目成本，提高决策效率，例如 Uddin 等[3]的一项研究展示了使用各种机

器学习算法（如 SVM、随机森林等）来探索项目成本超支的原因和频率的方法，并提出了数据驱动的成本超支情况预测系统。虽然当前研究为科研成本管理提供了重要洞见，但多数研究仅关注成本管理的宏观策略或针对特定领域的实施方式，缺少对科研过程中各个具体环节成本的详尽分析。特别是在成本密集的实验阶段，缺乏一种系统的评估与分析方法。为此，本研究提出了一种创新的科研设备成本评估框架，旨在自动化地从科研论文中抽取与实验成本相关的关键信息，根据设计的成本评估公式精准和高效地评估科研设备投入成本，这不仅填补了现有研究的空白，也为实验成本的精确评估和科研资源的有效利用提供了新的工具和方法。

本研究聚焦于计算机科学与物理学两个领域内的科研文献，利用 arXiv 数据库与 Paper With Code 网站所提供的论文数据构建了一个训练数据集，采用 LoRa 微调技术在基准模型 LLaMA2 上进行了微调，使其能够精确抽取计算机科学与物理学领域论文中关于实验设备与材料的详细信息，这一步骤为后续的成本分析奠定了坚实的基础。本研究根据平均情况分析的思想进一步设计了成本估算公式，对科研项目的设备成本进行了系统估算，通过在计算机视觉领域的实证分析，本研究发现论文平均研究成本呈指数级增长的趋势，同时不同模型架构和研究范式的性能提升存在客观上限。这些发现为科研项目的资源配置和预算管理提供了新的思路和依据，本研究的成果有望为科研管理者和决策者提供有价值的参考，推动科研资源的合理配置和科研事业的可持续发展。

## 1 相关研究

科研活动的成本评估是科研管理中的重要环节，需要综合考虑设备、材料、时间等多种要素，传统的成本评估方法难以有效应对科研活动的不确定性和创新性。近年来，人工智能技术的发展为科研成本评估提供了新的思路，通过自然语言处理等技术，可以从科技论文等非结构化文本中自动抽取与科研成本密切相关的信息，如材料、设备、实验参数等，并将其转化为结构化的知识表示，用于支撑科研项目的成本分析和管理。为此本节将从科研活动的成本评估方法和科技论文的信息抽取方法两个大方向对相关研究进行梳理。

### 1.1 科研活动的成本评估方法

科研活动的成本评估是科研管理中的重要环节，传统的科研活动成本评估方法主要包括专家评估法、类比估算法和参数估算法等 [4]。专家评估法依赖于专家的经验和判断，容易受到主观因素的影响；类比估算法通过与类似项目的对比来估算成本，但难以应对科研活动的创新性和不确定性；参数估算法建立了成本参数与影响因素之间的数学模型，但对参数的选择和量化存在挑战。这些方法往往依赖于手工收集和分析数据，难以适应科研活动日益增长的复杂性和数据量。

近年来，人工智能技术的发展为科研成本评估提供了新的思路。一些研究者尝试利用机器学习算法建立成本预测模型，通过对历史数据的训练来预测新项目的成本。Sajadfar 等 [5] 提出了一种结合特征工程和数据挖掘算法的成

doi:10.3772/j.issn.2095-915x.2024.05.009

本估算方法，利用线性回归和数据挖掘技术挖掘与企业资源计划（ERP）系统相关的制造过程数据，建立成本估算函数，该方法采用渐进式实施策略，提高成本估算的准确性和实用性；Liu 等 [6] 则基于支持向量回归（SVR）机器和粒子群优化（PSO）算法，建立了材料成本预测模型。该模型首先对实际数据进行预处理，然后利用 PSO 算法优化 SVR 的参数，进行数据挖掘和成本预测。研究表明，该模型能够很好地拟合实际材料成本数据，预测效果令人满意。然而，这些方法主要关注总体成本的预测，缺乏对科研活动中关键成本驱动因素的分析，如实验材料、设备等。此外，这些方法通常需要大量的历史项目数据作为训练样本，在一些新兴和交叉学科领域可能难以满足。

综上所述，现有的科研成本评估方法在应对日益复杂的科研活动时存在诸多局限，需要一种能够充分利用科技文献大数据、自动化获取关键成本信息、适应不同学科特点的新方法。近年来，随着自然语言处理技术的发展，特别是大型语言模型的出现，为材料设备信息的精准抽取提供了新的可能。相关技术如指令工程和高效微调使得大型语言模型能够在特定领域进行信息抽取，为材料设备成本的评估奠定坚实的数据基础。通过对科技文献的深入挖掘和分析，大型语言模型能够快速准确地识别和提取与研究成本密切相关的实验材料、设备等关键信息，并结合领域知识库实现成本的估算，从而为科研项目的成本评估和管理提供更加智能、高效的决策支持。本研究正是基于这一认识，提出了一种创新的基于大型语言模型的科研设备抽取和成本评估框架。

### 1.2 指令工程下的信息抽取技术

信息抽取任务（Information Extraction）是自然语言处理领域的重要任务之一，旨在从非结构化数据中抽取出结构化信息。传统的基于监督学习的信息抽取技术（如利用 CRF[7]、BERT[8] 等深度神经网络进行实体识别任务）大多遵循预训练 + 下游任务微调的范式，然而，这种传统方法的一个重要限制是对大量高质量标注数据的依赖，这不仅增加了成本，也限制了模型在面对复杂化数据时的适应性和扩展性。在论文材料设备信息抽取的实际任务场景中，往往缺乏大量高质量的训练数据，为此最新的信息抽取研究进一步探索利用指令工程（Prompt Engineering）、少样本学习（Few-shot Learning）等技术，在小样本、低资源场景下实现信息抽取，降低人工标注的成本。

随着 GPT-3.5[9] 模型的推出，开启了自然语言处理大模型元年，一年以来涌现出以 GPT-4[10]、Claude[11]、Gemini[12]、Qwen[13] 为代表的众多大模型，通过多任务训练和统一编码 [14] 的方式让模型展现出强大的语义理解能力和任务泛化能力，打破了自然语言任务之间的壁垒，将多种自然语言任务统一成序列到序列的生成任务，从而摆脱了对标注数据的依赖。指令工程指的是为大型语言模型设计和优化输入指令（或“提示”）的过程，旨在更有效率地借助模型的预训练知识库，完成特定的信息抽取任务。通过采用诸如思维链（Chain-of-Thought，CoT）[15]、语境学习（In-Context Learning）[16] 以及专家提示（Expert Prompting）[17] 等策略，研究人员能够引导模型进行深层次的语义分析，从而使之

精确理解特定信息抽取任务的需求，通过规范模型输出格式，使模型能高效地从非结构化文本中提取所需的结构化信息，而无需额外的标注数据集。指令工程不仅能将大型语言模型的泛用能力迁移到特定任务上，还可以提升其在特定信息抽取领域的性能和准确度。

这种对指令工程的深入研究和应用，已经开始在实际的信息抽取任务中展现其成效。Wei 等 [18] 将复杂的信息抽取任务转化为一个分两阶段进行的多轮问答问题，利用 ChatGPT 构建了一个名为 ChatIE 的零样本信息抽取多轮问答框架。首先，通过一次问答识别文本中的实体、关系或事件类型，然后利用链式抽取模板在多轮问答中进一步抽取与这些类型相关的信息，该方法在实体—关系三元组抽取、命名实体识别和事件抽取三个任务上的效果超过了一些全量训练的模型。Wang 等 [19] 提出了 InstructUIE 框架，利用自然语言指令引导大型语言模型完成信息提取任务，InstructUIE 在监督设置下与 BERT 的效果相当，在零样本设置下显著超过了 GPT-3.5 模型。Xiao 等 [20] 提出了一个名为 YAYI-UIE 的聊天增强型指令调整框架，结合对话数据和信息抽取数据进行训练，利用端到端的方法自动调整指令，使其适应不同的信息提取任务，该框架在中文数据集上达到了 SOTA 水平。

### 1.3 大模型高效微调技术

尽管提示工程通过设计精细框架和输入指令来引导大型语言模型执行特定任务，展现了大语言模型在零样本或少样本设置下处理信息抽取任务的能力。然而在本研究面临的实验材料与设备信息抽取任务中，目标信息通常包含大量专业术语、数值单位以及复杂的语义结构，对模型的领域知识理解和语义抽取能力提出了更高的要求，仅依赖提示工程很难使通用大模型在这一高度专业化的任务上达到理想的性能表现。如果采用特定领域数据对大模型直接进行微调，虽然可以让大模型充分学习到特定领域知识，提高大模型在领域任务上的能力，但是由于需要调整大量的参数，直接微调往往需要花费高昂的计算资源和可能持续数月的训练时间。为此我们需要采用一些高效微调的技术，如 Adapter Tuning[21] 或 LoRA（Low-Rank Adaptation）[22] 技术。Adapter Tuning 在模型的各层之间插入小型的可训练模块（称为 Adapter），只调整这些模块的参数而不改变原始预训练模型的权重，从而能够在保持预训练知识的同时，对特定任务进行有效的微调；LoRA 的核心原理基于对模型权重的低秩适配，这种方法能够在保持大部分预训练知识不变的同时，通过调整一小部分参数来实现对新任务的快速适应。

在信息抽取任务中，Jiao 等 [23] 结合 LoRA 技术提出了一个新颖的面向需求的信息提取框架 ODIE（On-Demand Information Extractor），这一框架允许模型根据用户指令生成结构化的信息表格，显著提升了在自动化评估和人类评估中提取表头和内容的准确性；Dagdelen 等 [24] 提出了一种从科学文本中提取结构化信息的创新方法，这项研究聚焦于材料科学领域，主要解决固态杂质掺杂、金属有机框架（MOFs）以及通用材料信息提取三项任务。通过使用少量标记数据对大语言模型进行高效微调，并以预定义的结构化模式（如 JSON 文档）输出信息，

doi:10.3772/j.issn.2095-915x.2024.05.009

展现了大模型通过微调展现出的将非结构化科学文本转换为结构化数据的潜力。

## 2 基于实验材料与设备信息的科研设备成本评估框架

在现代科研活动中，面对资源有限的现实，科研团队需要精确评估项目成本，确保资金的有效利用。传统的成本评估方法依赖于人工收集和分析数据，这不仅耗时耗力，而且容易受到主观判断的影响，难以应对科研活动中的复杂性和动态变化。尤其是在需要大量实验材料与设备的研究领域，如计算机科学和物理学，传统方法的局限性更为显著。这些领域的研究往往涉及高端设备和专业材料，其成本估算不仅需要考虑实验不同流程需要的不同材料设备，还需评估材料设备市场价格、设备运行时间等多个维度，这对成本评估的准确性和效率提出了更高要求。

针对当前科研设备成本评估方法研究的不足，本文从科研过程中重要的实验环节切入，设计了一个基于实验材料与设备信息的科研设备成本评估框架，如图 1 所示。该框架由两大核心模块构成：实验材料与设备信息抽取和

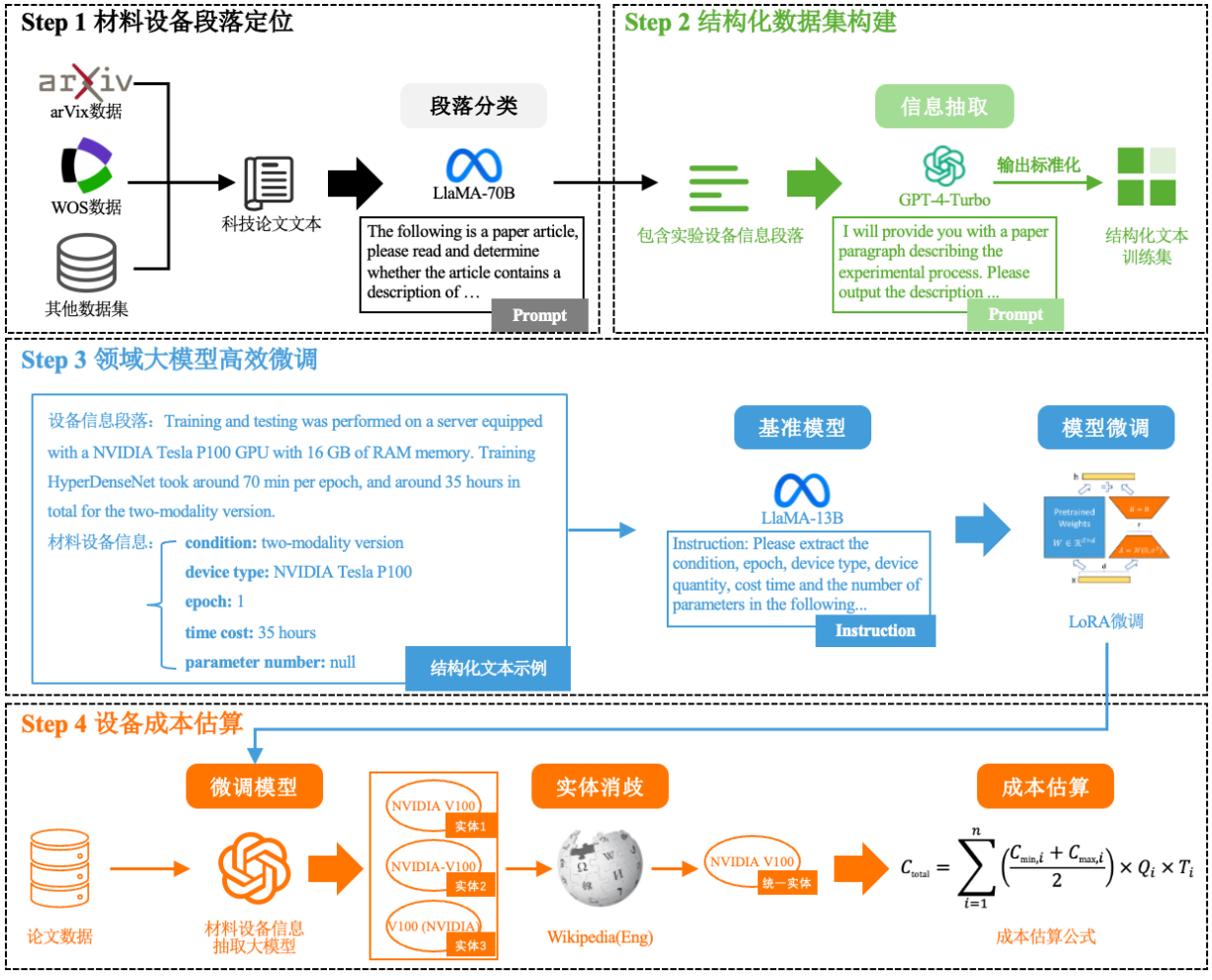


图 1 科研设备成本评估框架

成本评估。首先，实验材料与设备信息抽取模块以大型语言模型为基础，通过流水线策略的设计，实现从科学文献中自动定位并提取关键设备信息的功能。然后，成本评估模块通过对设备功能性与经济属性的分析，实现对多种类型材料与设备的统一价值评估，为科研项目成本提供更全面和精确的评估结果，帮助科研人员和项目管理者更好地规划和控制项目成本。

## 2.1 实验材料与设备信息抽取模块

实验材料与设备信息抽取模块采用“粗筛选＋精解析”的设计策略，先从海量的文本中定位包含材料与设备信息的潜在段落，然后再从筛选出的文本中进行关键信息的提取，以保证整个信息抽取模块的准确性和效率，为后续的成本评估提供精准的基础信息。

### 2.1.1 实验材料与设备信息定位

为了尽可能召回包含实验材料与设备信息的文本段落，本研究设计了两种互补的信息定位策略来引导大型语言模型完成该目标。第一种策略采用“粗召回—分类”的两阶段方法实现对成本段落的快速定位。首先，使用正则表达式编写关键词模板，通过匹配段落中所有与材料、设备和成本相关的关键词，包括计算设备如“GPU”“NVIDIA”等，时间成本如“hours”“days”等，从而尽可能多地召回目标句；随后，设计用于文本分类任务的指令，进一步检验通过模板匹配采集到的句子是否满足需求。该指令如图 2 所示，由任务指令（红色）、任务描述（蓝色）以及输出指令（绿色）三个部分组成，保证模型在理解任务需求的同时能够以固定的格式输出结果。

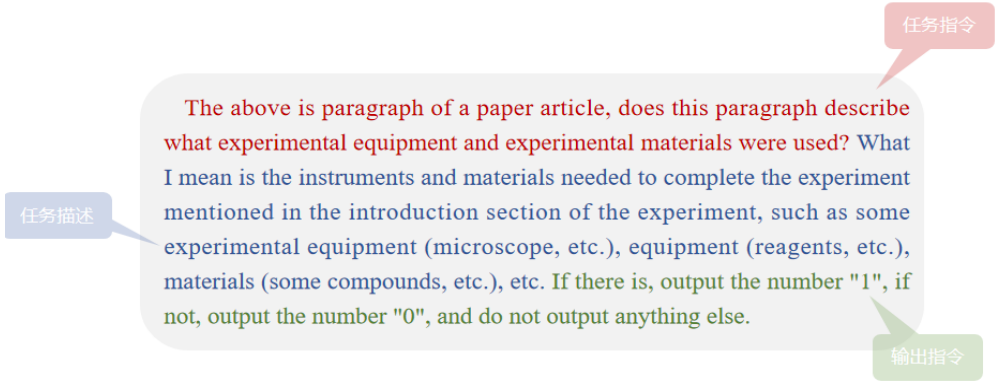


图 2 实验材料与设备信息定位指令模版

第二种策略则面向难以通过既定模板匹配出的句子，如一些具有特殊用途的设备（X-ray）和材料等，在这些未匹配出的文本上进一步进行分类操作。在该策略中，为了尽可能降低误检率，在模型输出结果上作出了更加严格的约束，假设模型在分类句子时生成正例结果“1”的概率为 α，当 α 大于 0.8 时才将检测出的句子视为可信。

### 2.1.2 实验材料与设备信息识别

在完成目标句的筛选后，本研究采样了少量的句子用于训练数据的构建，完成材料与设备信息抽取模型的设计，实现批量快速的信息抽取，服务于研究最终的成本评价目标。详细的训练数据获取过程以及模型构建过程如下。

doi:10.3772/j.issn.2095-915x.2024.05.009

（1）材料与设备信息识别数据获取

不同学科领域在实验过程中使用的材料和设备会有较大的差异。以本研究关注的计算机领域与物理学领域为例，计算机领域的实验成本主要来源于计算资源的数量、性能及其在训练特定参数规模模型时的轮次和耗时，而对于物理学领域，不同材料都有其固定的成本开销，同时在实验时间上也有一定的耗费，但这一信息较少在文中提及，因此在本研究中没有考虑。不同类型材料和设备需要考虑的成本属性如表 1 所示。

**表 1　实验材料和设备的成本属性**

| | 物理材料与设备 | 计算机设备 |
|---|---|---|
| 成本属性 | 名字（Name） | 设备类型（Device type） |
| | 描述（Description） | 设备数量（Device quantity） |
| | 数量（Quantity） | 轮次（Epoch） |
| | — | 时间开销（Time cost） |

基于对材料和设备重要成本属性的分析，本研究仍采用指令工程技术来引导性能更佳的大型语言模型来对采样的目标句进行结构化解析，如表 2 所示。

**表 2　材料与设备信息识别数据获取示例**

| 步骤 | 角色 | 文本 |
|---|---|---|
| 指令 | SYSTEM | Now you need to help me complete the information extraction task. |
| | USER | I will provide you with a paragraph describing the training cost of model in some condition. Please out put the condition (fully described),epoch ,device type ,device quantity and cost time of each condition in the list of JSON documents like [{}, {}], If a key does not exist, output value as None. You should pay attention to the model I gave you, don’t output the cost of other model |
| 输入 | USER | We use 8 NVIDIA A100 GPUs for model training, and each training is finished within 3 days for large models. During training, we apply SpecAug  using 2 frequency-masks with parameter $F$ of 27, and 10 time-masks with parameters $p$ of 0.05. |
| 输出 | ASSISTANT | [{<br>“condition”：“default”，<br>“epoch”:None,<br>“device_type”：“Nvidia A100”，<br>“device_quantity”: 8,<br>“time_cost”：“3 days”，<br>“parameter_number”:None<br>}] |

由于大型语言模型性能的局限，解析出的结果会出现一定错误，为了保证训练数据的质量，本研究在模型识别的结果上开展了人工验证。

（2）模型构建

在构建好的训练数据的基础上，本研究采用指令微调的方式来引导通用的大型语言模型在材料和设备信息抽取任务上取得更好的应用效果。具体地，本研究采用了 LORA 微调策略，即在原有的模型上添加小规模的参数来学习任务信息。采用该方法主要原因在于计算资源以及训练数据数量的局限性，以及该方法在类似任务上取得的成功应用。

假设通过大型语言模型的总参数量为 $W_{base}$，LORA 微调策略的目标是在此基础上添加少量参数 $\Delta W$，而这一新添加的参数由多个升维矩阵 $A$ 和降维矩阵 $B$ 组成，由此：

$$W_{base}+\Delta W= W_{base} + BA \quad (1)$$

若 $W_{base} \in R^{d\times k}$，则 $B \in R^{d\times r}, A \in R^{r\times k}$，其中 $r$ 远小于 $d$ 和 $k$，因此 $\Delta W$ 要远小于 $W_{base}$，因此使用少量计算资源来微调大规模参数模型成为可能。本研究在 $W_{base}$ 的每一个键—值线性层的旁路都添加了上述可学习的参数，可学习参数的比例占模型全参数比例的 0.0503%。最后，冻结 $W_{base}$，将抽取实验材料与设备成本属性的指令作为输入，Json 输出结果作为输出对 $\Delta W$ 进行微调。

## 2.2 成本评估模块

利用微调后的大模型对科学论文进行材料设备信息的结构化抽取后，本研究通过链接消歧—价格估算的方式对实验成本进行计算。对于抽取出来的实体先通过 Wikipedia（English）进行查询，如果不同实体指向了相同的 Wikipedia 信息页面，本研究将其视为是同一实体的不同展现形式，并对它们进行合并以消除歧义，然后通过式（2）对合并后的材料设备信息实体进行成本估算，得到最后结果。

### 2.2.1 实体链接

在科学文献中，对同一材料或设备的描述可能采用多种不同的表达方式。例如，“综合物性测量系统”既可以被完整地表述为“Physical Property Measurement System”，也可以简略地称作“PPMS”，此外还存在其他形式的实体歧义，如大小写差异、是否包含注释等，如表 3 所示。

表 3 设备名称歧义示例

| 情况 | 示例 |
|---|---|
| 设备名称存在大小写区分 | {entity: OSCAR, mention: Oscar} |
| 设备名称简写 | {entity: Physical Property Measurement System, mention: PPMS}<br>{entity: Chemical Vapor Deposition System, mention: Chemical Vapor Deposition (CVD) system } |
| 设备名称包含符号 | {entity: x-ray diffractometer, mention: x ray diffractometer },<br>{entity: light scattering measurement equipment, mention: measurement equipment (light scattering)} |
| 设备名称包含注释 | {entity: Microprobe, mention: Microprobe (2015)} |

为了精确和高效地估算成本，本研究采纳了基于 Wikipedia 的实体链接方法来解决实体的歧义问题。作为一个庞大的在线知识库，Wikipedia 涵盖了从科技术语到实验材料等广泛的主题，几乎所有的材料和设备实体都能在其数据库中找到相应的条目。此外，Wikipedia 的数据高度结构化，包括信息框（infoboxes）、分类（categories）、重定向页面（redirect pages）以及消歧义页面（disambiguation pages）等元素，这些结构化的信息极大地提升了实体识别和消歧的精准度。通过利用 Wikipedia-based Entity Linking[25] 方法，本研究能够有效地识别和统一这些多样化的表述，从而在成本估算中实现更高的准确性和效率。

### 2.2.2 成本估算

精确计算科学实验中材料与设备的成本是一个高度复杂的任务。即使是同一名称的设备或材料，其价格也可能因品牌、型号等多种因素而产生显著波动，这种价格波动给预算制定和成本控制带来了不确定性。为了应对这一挑战，本研究借鉴了计算机科学领域中平均情况分析（Average-Case Analysis）的概念。平均情

doi:10.3772/j.issn.2095-915x.2024.05.009

况分析是一种算法性能评估方法，不同于最坏情况分析（Worst-Case Analysis）和最佳情况分析（Best-Case Analysis），它关注的是算法在所有可能的输入上的平均表现。这种分析方法提供了一种更加贴近现实的性能预测，因为该方法考虑了算法在正常各种输入情况下的性能，而不仅仅是在极端输入情况下的表现。

将平均情况分析的核心理念应用到科学实验材料与设备的成本估算中，意味着在计算成本时将考虑到价格波动的整体分布，而不仅仅是最高价或最低价，从而提供一个更加准确的成本估算。为此，本研究提出了如下成本估算公式。

$$C_{\text{total}} = \sum_{i=1}^{n} \left( \frac{C_{\min,i} + C_{\max,i}}{2} \right) \times Q_i \times T_i \quad （2）$$

假设一个实验需要 $n$ 项材料或者设备完成，其中 $Q_i$ 为在实验中第 $i$ 项材料或设备的数量；$T_i$ 为使用第 $i$ 项材料或设备的实验时长（如果使用），如果某项材料或者设备的使用成本不涉及时间长度，则 $T_i$ 可从公式中省略；$C_{\min,i}$ 和 $C_{\max,i}$ 分别代表第 $i$ 项材料或设备的最小和最大成本。

此公式旨在综合考虑同一材料或设备在不同情况下的价格波动，通过计算其平均成本来提供一个既实用又具有代表性的成本估算。这种方法不仅允许本研究在面对价格信息不完全或变化大的情况下进行有效的成本规划，而且还提高了成本估算的准确度和可靠性。

## 3 实验结果和分析

### 3.1 实验设置

（1）实验数据

本研究所依托的原始数据集经由 Paper With Code 和 arXiv 数据库获得，覆盖计算机和物理学两大领域，涉及自 2015—2023 年发表的学术论文总计 11319 篇，其中计算机领域论文 3192 篇，物理学领域 8127 篇。基于此原始数据集，本研究进一步采用基于大型语言模型的预处理技术，筛选出含有实验材料与设备信息的段落共 5228 条。此外，通过精心设计的指令模板，本研究构建了包含 5228 条数据的数据集，同时该数据集进一步细分为训练集 4928 条、验证集 100 条以及测试集 200 条。

（2）模型选择

在构建科研设备成本评估框架的过程中，针对实验材料与设备信息的准确定位、识别及训练数据的收集，本研究采纳了三种差异化的计算模型以实施相应任务。首先，为了筛选含有材料及设备信息的描述性文段，本研究使用了具有良好语言理解能力的 70B 参数 LLaMA2 模型。其次，通过信息抽取提示模板的设计，利用性能更卓越的 GPT-4-turbo 模型对文段进行精细化解析，旨在构建一个高质量的训练数据集。最后，本研究采用了一个规模较小（13B 参数）的 LLaMA2 模型进行模型微调优化，实现将大规模参数模型的信息抽取能力迁移到较小规模模型上，从而提高实验材料与设备信息的识别效率。

（3）参数设置

训练模型的参数设置详见表 4，训练批次大小为 8，训练轮次为 10，学习率为 1e-4。

（4）实验设备

为完成实验材料与设备信息抽取以及成本评估等关键研究内容，本研究采用了 2 块 A100 GPU 来执行大模型的提示工程与指令微调工作，具体的实验设备与环境如表 5 所示。

表 4　大型语言模型微调参数设置

| 参数名称 | 参数含义 | 参数设置 |
|---|---|---|
| batch_size_training | 训练批次大小 | 8 |
| num_epochs | 训练轮次 | 10 |
| lr | 学习率 | 1e-4 |
| gamma | 伽马值 | 0.85 |
| val_batch_size | 验证批次大小 | 1 |
| peft_method | 调优框架 | lora |

表 5　实验设备与环境

| 实验环境 | 环境配置 |
|---|---|
| 操作系统 | Ubuntu 22.04 LTS |
| GPU | 2 × NVIDIA A100 SXM4 80GB |
| 内存 | 1.5 T |
| 编译器 | Python 3.10.11 |
| 深度学习框架 | Pytorch 2.0.0 |

## 3.2　评价指标

本研究为了验证基于实验材料与设备信息的科研设备成本评估框架的可靠性，将首先着重对实验材料与设备信息抽取的性能进行评价，使用信息抽取研究通用的准确率（precision）、召回率（recall）和 F1 值进行评测。

## 3.3　实验结果与分析

### 3.3.1　材料设备信息抽取性能评估

本研究选择了 LLaMA2-13b-base、GPT-3.5-turbo-0125 和 LLaMA2-13b-lora（本实验调优后的模型）三个大语言模型在测试集上进行实验，性能评估覆盖了计算机科学和物理学两个领域，重点关注从各自领域的科技论文中抽取材料和设备细节的能力，实验结果如表 6 所示。

实验结果表明，在进行计算机科学和物理学领域内的材料设备信息抽取任务时，LLaMA2-13b 的基础模型表现较为一般，这在某种程度上反映了该基础模型在理解这两个领域知识的深度与广度方面存在的局限。然而，通过使用 LoRA 技术进行高效调优，LLaMA2-13b-lora 模型在这两个领域内展现出显著的性能提升，部分性能指标甚至超越了 OpenAI 推出的 GPT-3.5-turbo-0125 模型。这一成果突显了 LoRA 技术在特定应用领域调优大型语言模型的潜力，尤其是在精确抽取材料设备方面的应用，并为后续的成本评估工作奠定了坚实的基础。

表 6　实验结果

| 研究领域 | 计算机 | | | | 物理 | | | |
|---|---|---|---|---|---|---|---|---|
| 评价指标 | P | R | Macro F1 | Micro F1 | P | R | Macro F1 | Micro F1 |
| LLaMA2-13b-base | 0.3871 | 0.4953 | 0.4038 | 0.4345 | 0.5335 | 0.5929 | 0.5437 | 0.5725 |
| GPT-3.5-turbo-0125 | 0.5042 | 0.6453 | 0.5017 | 0.5661 | 0.6568 | 0.7035 | 0.6487 | 0.6793 |
| LLaMA2-13b-lora | 0.6221 | 0.5966 | 0.4939 | 0.6091 | 0.7284 | 0.6679 | 0.6890 | 0.6969 |

### 3.3.2　科研设备平均成本变化分析

本研究选择了 Paper With Code 网站计算机视觉领域下的高 star（收藏数大于 10000）科技论文 292 篇作为典型样例进行小样本成本估算。这一策略确保了样本论文在学术与工业界的广泛认可度和应用价值，反映出其影响力和代表性。

如图 3 所示，本研究按照时间顺序进行了

doi:10.3772/j.issn.2095-915x.2024.05.009

论文数量和平均论文成本的初步统计。数据表明，无论是典型论文数量还是平均论文成本，都随着年份的增加而呈现上升趋势，特别是在2020年，可以明显看到这一趋势的加速。本研究注意到，这与CV界的Transformer架构ViT[26]的提出时间相吻合。自ViT的提出以来，计算机视觉领域的研究开始更多地采用Transformer架构，而这一架构模型训练所需的数据量也通常比之前的模型多，这就导致科技论文实验需要更高性能的计算资源，成本开销大幅度提升，这与本研究的成本估算的实验结果也保持一致。

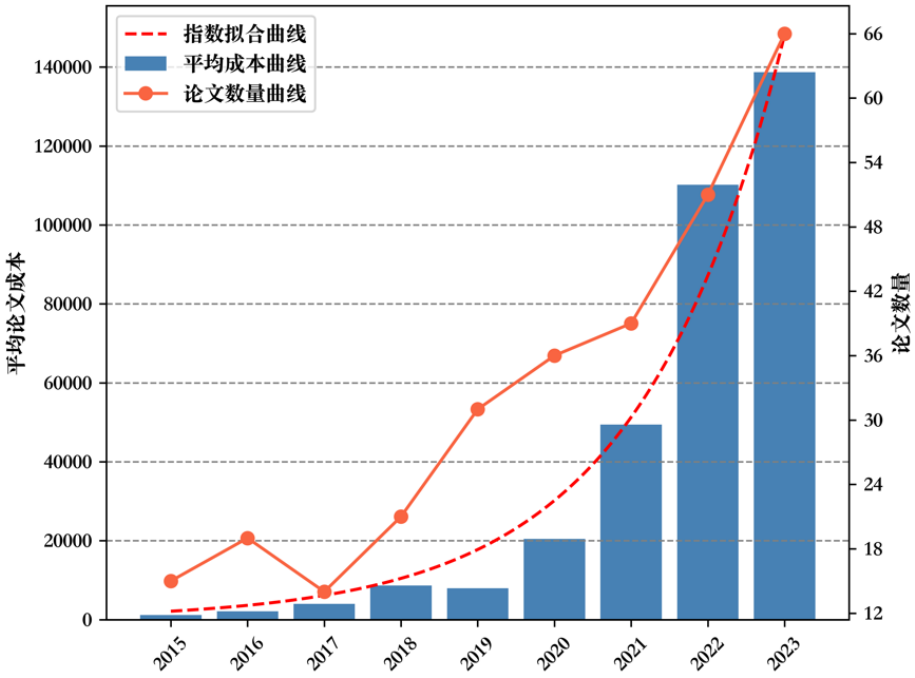


图3 样例成本估算结果

OpenAI于2022年末发布的ChatGPT语言模型，更是引发了学界对大模型研究的高度关注，这进一步推动了2023年计算机视觉领域平均论文成本的增长。综合以上分析可以看出，在当前大模型主导的研究范式下，论文的平均研究成本正呈现出指数级的增长态势，这表明计算资源已经成为制约计算机视觉领域科研产出的关键因素之一。

### 3.3.3 科研设备投入与模型性能关联分析

为更细致地刻画科研设备对关键任务性能的影响，本小节聚焦目标检测这一计算机视觉的核心任务，对历年SOTA模型的成本效益进行了实证分析。本小节选取了Paper With Code网站中目标检测任务在COCO test-dev数据集上的SOTA模型对应的论文作为研究对象，对其进行了成本估算分析。图4展示了不同时期的SOTA模型在COCO数据集上的性能表现（BOX mAP分数）与其对应论文的研究成本之间的关系。散点图中每个散点代表一个SOTA模型，其中散点的颜色代表模型所采用的基本架构。

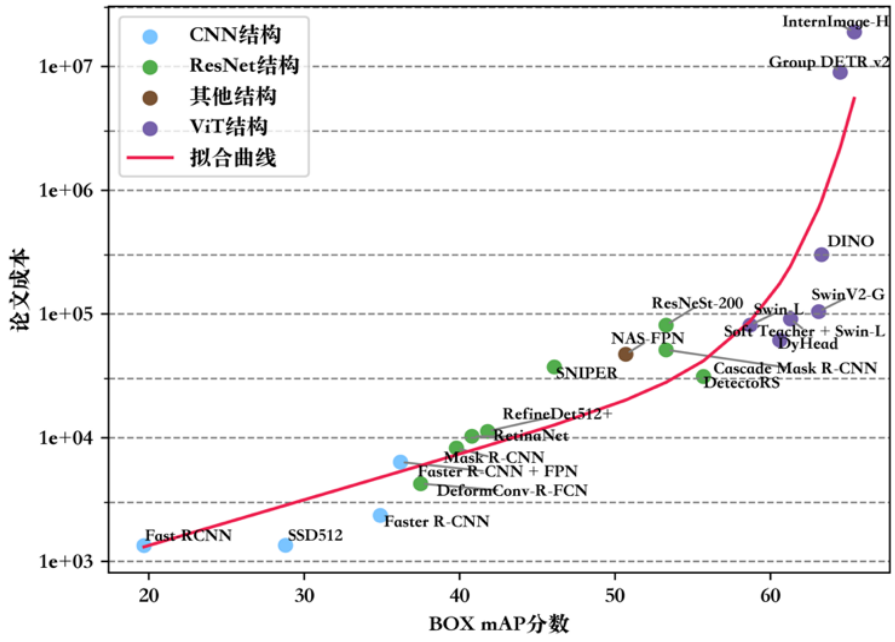


图 4　模型效果与设备关系散点图

如图 4 所示，随着科研设备成本的增加，目标检测模型的性能整体上呈现出上升的趋势。这表明在目标检测任务上，投入更多的计算资源和设备成本，通常能够带来性能的提升，这一点与直觉相符，也与前文分析论文设备投入整体成本变化趋势的结论一致；同时，本研究也注意到不同模型架构的性能表现和成本效益存在明显差异。以传统的 CNN 结构的模型为例，其性能随着成本的增加而提升，但当达到一定程度后，继续增加设备成本带来的性能提升变得非常有限；相比之下，以 ResNet 为代表的更先进的模型架构，在相同设备成本下能够取得更优的性能表现，这得益于残差网络更强大的特征提取和表示能力，使其能够更好地捕捉和建模图像中的关键信息。ResNet 结构的出现，在一定程度上突破了传统 CNN 模型的性能瓶颈，通过构建更深、更复杂的网络，可以进一步推动目标检测任务的性能提升。然而，ResNet 结构的性能提升也并非没有上限，当模型复杂度和计算成本达到一定程度后，其性能增益同样会趋于饱和。

近些年来，ViT 架构的模型为目标检测任务的发展注入了新的活力，其强大的特征学习能力使其在目标检测任务上取得了突破性的进展。从图中可以看出，ViT 模型的出现进一步推高了目标检测任务的性能上限，但与此同时，其对计算资源的需求也大幅增加。为了充分发挥 ViT 模型的性能优势，需要在海量数据上进行大规模的模型训练，这无疑会带来更高的设备投入成本。同时本研究注意到，早期的深度学习模型能力较为单一，往往只作用于单一数

doi:10.3772/j.issn.2095-915x.2024.05.009

据集的单一任务上；而现在的大模型一方面具备处理在多个数据集上进行多种任务的能力，另一方面大模型的评测往往需要大量的对比消融实验，这也进一步加剧了科研设备成本呈指数型增长的现象。

综上所述，对于特定的模型架构或研究范式，客观上存在一个性能上限，当设备投入成本和模型复杂度超过某个临界点后，继续增加设备投入所带来的性能提升将变得非常有限，此时设备投入的边际效益将显著降低。这对于科研人员和项目管理者优化资源配置、控制成本风险具有重要的指导意义，通过及时发现和预警低效益或高风险的研究方向，可以更加科学地规划和调整项目预算，提高科研投入的效益。

## 4 讨论与结论

本研究旨在实现对物理和计算机领域学术论文中的材料设备信息进行低成本、高效率及高质量抽取，为此，本研究针对不同的学科领域设计了特定的指令集，通过这些指令集引导大型语言模型深入挖掘不同领域学术论文对于材料设备描述的特征，同时为不同领域论文中材料设备信息的抽取设计了专门的抽取范式。本研究构建了材料设备信息抽取任务的训练数据集，利用低秩微调技术对大模型进行精细调优，这种方法显著提高了大模型在特定领域内对材料和设备信息进行结构化抽取的能力，为后续的材料设备成本计算工作提供了坚实的基础，进而为科研人员和项目管理者在资源规划和成本估算方面提供有力的支持。

本研究通过对计算机视觉领域的实证分析，验证了基于大型语言模型的成本评估方法的有效性。研究结果表明，本文提出的框架能够准确估算不同时期论文的平均研究成本，揭示出计算资源对科研产出的重要影响。同时，通过对目标检测任务的成本效益分析，本研究还发现，不同模型架构的性能提升和设备投入之间存在一个客观的临界点，超过这一临界点后，继续增加投入的边际效益将显著降低。这些发现与实际科研活动的规律相吻合，证明了本文提出的成本评估方法能够为科研项目的资源配置和风险管理提供有价值的决策参考。尽管当前的分析还主要集中在计算机视觉领域，但本研究提出的方法论具有一定的普适性，未来可以进一步拓展到更多学科和任务场景中，为科研管理提供更加全面和精准的成本评估工具，全面提升科研项目管理的质量和效率。

## 参 考 文 献

[1] 罗准辰，赵赫，叶宇铭，等．基于文献链接信息分析的科技资源风险评估 [J]. 中文信息学报，2020, 34(5): 64-73.

[2] KAIL E, KACSUK P, KOZLOVSZKY M. New aspect of investigating fault sensitivity of scientific workflows[C]//2015 IEEE 19th International Conference on Intelligent Engineering Systems (INES). Bratislava, Slovakia: IEEE, 2015: 185-188.

[3] UDDIN S, ONG S, LU H. Machine learning in project analytics: a data-driven framework and case

study[J]. Scientific Reports, 2022, 12(1): 15252.

[4] LOCK D. Cost estimating[C]//In The Essentials of Project Management. Routledge, London, 2001: 31.

[5] SAJADFAR N, MA Y. A hybrid cost estimation framework based on feature-oriented data mining approach[J]. Advanced Engineering Informatics, 2015, 29(3): 633-647.

[6] LIU S, QI G, ZHEN L, et al. Research on the Prediction Model of Material Cost Based on Data Mining[J]. The Open Mechanical Engineering Journal, 2015, 9(1): 1062-1066.

[7] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[8] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805, 2019.

[9] BROWN T, MANN B, RYDER N, et al. Language Models are Few-Shot Learners[C]//Advances in Neural Information Processing Systems: Vol. 33. Curran Associates, Inc., 2020: 1877-1901.

[10] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.

[11] WU S, KOO M, BLUM L, et al. A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology[J]. arXiv preprint arXiv:2308.04709, 2023.

[12] ANIL R, BORGEAUD S, WU Y, et al. Gemini: A family of highly capable multimodal models[J]. arXiv preprint arXiv:2312.11805, 2023.

[13] BAI J, BAI S, CHU Y, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.

[14] MISHRA S, KHASHABI D, BARAL C, et al. Cross-Task Generalization via Natural Language Crowdsourcing Instructions[C]//Muresan S, Nakov P, Villavicencio A. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 3470-3487.

[15] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[16] DONG Q, LI L, DAI D, et al. A survey on in-context learning[J]. arXiv preprint arXiv:2301.00234, 2022.

[17] XU B, YANG A, LIN J, et al. Expertprompting: Instructing large language models to be distinguished experts[J]. arXiv preprint arXiv:2305.14688, 2023.

[18] WEI X, CUI X, CHENG N, et al. Zero-shot information extraction via chatting with chatgpt[J]. arXiv preprint arXiv:2302.10205, 2023.

[19] WANG X, ZHOU W, ZU C, et al. Instructuie: Multi-task instruction tuning for unified information extraction[J]. arXiv preprint arXiv:2304.08085, 2023.

[20] XIAO X, WANG Y, XU N, et al. YAYI-UIE: A Chat-Enhanced Instruction Tuning Framework for Universal Information Extraction[J]. arXiv preprint arXiv:2312.15548, 2023.

[21] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]// International conference on machine learning. PMLR, 2019: 2790-2799.

[22] HU E J, SHEN Y, WALLIS P, et al. Lora: Low-

doi:10.3772/j.issn.2095-915x.2024.05.009

rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.

[23] JIAO Y, ZHONG M, LI S, et al. Instruct and Extract: Instruction Tuning for On-Demand Information Extraction[C]//2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023. Association for Computational Linguistics (ACL), 2023: 10030-10051.

[24] DAGDELEN J, DUNN A, LEE S, et al. Structured information extraction from scientific text with large language models[J]. Nature Communications, 2024, 15(1): 1418.

[25] HACHEY B, RADFORD W, NOTHMAN J, et al. Evaluating Entity Linking with Wikipedia[J]. Artificial Intelligence, 2013, 194: 130-150.

[26] DOSOVITSKIY A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.