使用 NLP 技术和基于 LLM 的检索增强技术进行自动文献综述

一代

努沙特·法提·阿里
计算机科学与工程系
军事科学技术学院
孟加拉国达卡
nurshatfateh@gmail.com

穆罕默德·马赫迪·穆赫塔西姆
计算机科学与工程系
军事科学技术学院
孟加拉国达卡
mahdimohtasim@gmail.com

沙基尔·莫沙罗夫
计算机科学与工程系
军事科学技术学院
孟加拉国达卡
shakilmrf8@gmail.com

T. 戈皮·克里希纳
计算机科学与工程系
军事科学技术学院
孟加拉国达卡
gopi.mistbd@gmail.com

抽象的**本研究提出并比较了多种使用自然语言处理 (NLP) 技术以及基于大型语言模型 (LLM) 的检索增强生成 (RAG) 自动生成文献综述的方法。不断增长的研究论文数量对手动文献综述提出了巨大的挑战，也导致了对自动化的需求不断增长。本研究的主要目标是开发一个能够仅根据 PDF 文件作为输入自动生成文献综述的系统。为了满足主要目标，我们评估了几种自然语言处理 (NLP) 策略的有效性，例如基于频率的方法 (spaCy)、Transformer 模型 (Simple T5) 以及基于大型语言模型的检索增强生成 (RAG) (GPT-3.5-turbo)。本研究实验选择了 SciTLDR 数据集，并采用三种不同的技术实现了三个不同的文献综述自动生成系统。ROUGE 分数用于评估这三个系统。根据评估，大型语言模型 GPT-3.5-turbo 获得了最高的 ROUGE-1 分数，为 0.364。Transformer 模型排名第二，spaCy 排名最后。最后，我们为基于大型语言模型的最佳系统创建了图形用户界面。**

索引术语**—T5、SpaCy、大型语言模型、GPT、ROUGE、文献综述、自然语言处理、检索增强生成。**

我简介

文献综述对学者来说已变得非常重要。它为研究人员提供了特定领域先前研究成果的全面概述，并帮助学者发现过去理解中的不足之处。它有助于开展未来的研究，并让研究人员了解哪些领域可以提供重要的投入。然而，进行文献综述可能非常繁琐，因为需要阅读的文献实在太多。由于发表的研究论文数量庞大，审查所有相关研究并提取检索相关信息可能是一项耗时、繁琐且容易出错的任务。由于这些困难，人们对文献综述流程的自动化越来越感兴趣 [1]。自动化系统可以使用自然语言处理技术和机器学习算法来分析大量文本，提取相关细节，并创建结构化的摘要[2]。

本研究的主要目标是开发一个系统，该系统仅使用相关论文的 PDF 文件作为输入，即可自动生成研究论文的文献综述部分。我们实现了几种自然语言处理技术，例如基于频率的方法、基于 Transformer 的方法和基于大型语言模型的方法，并进行了比较，以找到最佳方法。本研究选择了 SciTLDR 数据集 [3]。第一个过程使用基于频率的方法。这里使用了名为 spaCy [4] 的库。第二个过程使用基于 Transformer 的模型。这里使用了简单的 T5 模型。最后一个过程基于使用大型语言模型。这里使用了 GPT-3.5-TURBO-0125 模型。使用 ROUGE 分数 [5] 进行评估和比较。然后确定最佳方法并创建基于图形用户界面的工具。

文献综述流程的自动化可以帮助学者节省时间，专注于与其研究最相关的文章。它还可以减少综述过程中出现错误或偏见的可能性。本文的重点如下：

- 所有三种考虑的 NLP 方法（例如 spaCy、T5 和 GPT-3.5-TURBO-0125 模型）都可以在自动生成文献综述方面产生令人满意的结果。
- 基于 LLM 的模型在生成文献评论方面优于 T5 和 spaCy。

II. L文献R回顾

Silva 等人 [6] 提出了一个自动生成系统文献综述的框架。他们专注于四个技术步骤：检索、筛选、映射和综合。针对特定问题，研究人员会进行广泛的检索，以尽可能多地查找相关研究，包括查阅参考文献列表、搜索互联网数据库以及查阅已发表的文献。筛选通过将检索范围限制为与特定综述相关的论文来缩小搜索范围，旨在突出可能影响政策的重要发现和事实。映射用于理解特定领域的研究活动，吸引利益相关者参与，并确定与综述重点相关的优先事项。综合整合来自众多来源的数据，并提供研究结果的概述。研究问题的制定、报告阶段和同行评审也是撰写系统文献综述的一些步骤。

随着科学的快速发展，同行评审的出版物数量呈指数级增长。因此，袁等人 [7] 探索了利用机器学习技术、自然语言生成、多文档摘要和多目标优化来实现科学评审的自动化。他们探讨了综合评审的生成，并指出了建设性反馈与人工撰写的评审相比的局限性。本研究中使用的模型尚无法完全实现文献评审的自动化，需要人工评审。

Karakan 等人 [8] 对现有的系统性文献综述工具进行了全面分析。他们探索了在综述流程各个阶段实现自动化的潜力，强调需要设计一个整体的工具来有效应对研究人员面临的挑战。他们讨论了两种完成研究的方法：快速综述和半结构化访谈。快速综述强调解决软件工程师在日常工作中遇到的问题、困难和挑战的决策程序。半结构化访谈用于

探索研究人员的经验、挑战、策略、系统文献综述工具的优势、劣势以及软件工程有效支持的要求。

Jaspers 等人 [9] 专注于利用机器学习技术实现文献综述和系统评价的自动化。他们概述了不同机器学习技术的优缺点，并详细讨论了文献综述的自动化流程。但该论文缺乏跨领域的实践验证和详细的说明。

景点。

Tauchert 等人 [10] 对自动化文献综述进行了简要概述。他们强调了自动化在系统综述流程各个阶段的潜力。该论文探讨了整合计算技术以简化检索、筛选、提取和合成等任务的重要性。论文也承认需要进一步研究以应对挑战。

并提高自动化方法的有效性。

Tsai 等人 [11] 对自动文献综述工具进行了简要概述。他们探讨了该领域的现有研究、手动进行文献综述所面临的挑战以及自动化流程的潜在优势。他们的研究重点是评估 Mistral LLM 在学术研究领域的有效性。

Susnjak 等人 [12] 探讨了系统文献综述 (SLR) 与 LLM 交叉领域的差距。他们还强调了应对研究综合阶段挑战的必要性，并强调了利用数据集微调 LLM 以提高知识综合准确性的潜力。本研究旨在通过提出一个系统文献综述自动化框架来弥合这一差距。

已讨论的大多数相关研究主要集中在使用 NLP 技术和 LLM 自动化文献综述过程的潜力和挑战上。它们都没有提出一个完整的系统流程，让用户能够仅使用 PDF 和 DOI 直接生成文献综述。相比之下，本文提出并实现了三个独特的端到端文献综述自动化系统流程和程序。这项研究工作还促成了一个 UI 工具的实现，用户可以直接上传 PDF 并自动生成文献综述片段，而无需任何额外工作。此外，本文还使用 ROUGE 分数对基于频率的方法、基于变换器的方法和基于 rag 的方法等不同方法进行了比较分析，这有助于发现这些方法在此任务中的有效性。

III. S系统D设计

研究分为四个阶段进行：1. 确定研究目标。2. 提出多种自动生成文献综述的程序。3. 评估多种程序以找到最佳方法。4. 最终系统开发。

A.数据集选择

本研究工作选择了 Hugging Face 的 SciTLDR 数据集 [13]。该数据集包含科学文献的摘要，包含 5400 个 TLDR，这些 TLDR 源自 3200 多篇论文。它包含作者撰写的和专家生成的科学文献 TLDR。精选研究文章的摘要、引言和结论 (AIC) 或论文全文作为“来源”，相应文章的摘要作为“目标”。提出的三个程序都只使用了这两个属性。spaCy 方法无需训练，但该数据集可用于测试目的。T5 模型使用 SciTLDR 数据集进行基于 Transformer 的方法训练，然后在测试数据集上进行评估。对于基于 LLM 的方法，该数据集用作模型的知识库。

B. 使用 spaCy 的基于频率的方法

第一个过程利用了 spaCy 的基于频率的方法。首要任务是构建模型管道。该模型管道以文本作为输入，并使用 spaCy 库将文本转换为 NLP 标记。然后，通过删除停用词和标点符号来完成预处理步骤。之后，计算每个单词的词频，这有助于计算单个句子的权重。句子权重代表该句子的重要性。最后，选择排名前 10% 的句子作为最终输出。之后，使用 ROUGE 分数对模型进行评估，以了解其性能概况。spaCy 模型的概览如图 1 所示。

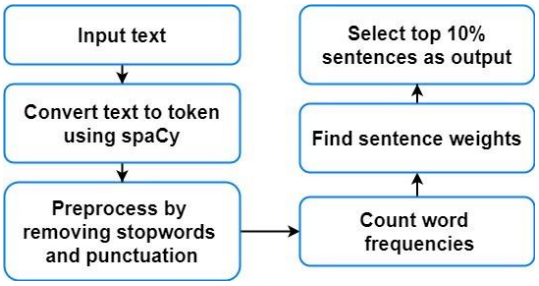

图 1：**构建 spaCy 模型**

下一步是使用 spaCy 模型实现系统流程，自动生成文献综述片段。该系统以多篇论文的 DOI 和 PDF 文件作为输入。它使用 Requests 库从 DOI 中收集论文标题和第一作者姓名。然后，它使用 PYPDF2 和正则表达式 (RE) 库仅收集每个 PDF 的结论部分。之后，它使用之前实现的 spaCy 模型获取每篇论文的摘要。之后，它会进行后处理并合并所有摘要，以生成连贯的文献综述片段。spaCy 模型的系统流程概览如图 2 所示。

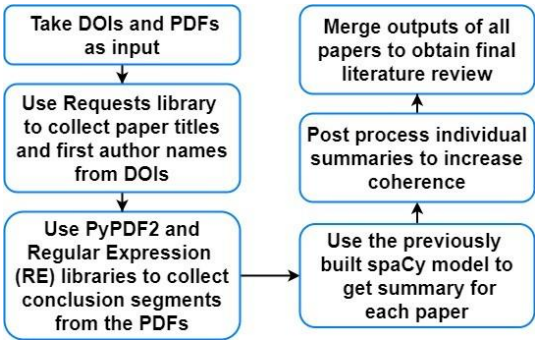

图2：**使用 spaCy 的管道**

C. 利用基于 Transformer 的 T5 模型的程序

第二种方法利用基于 Transformer 的 Simple T5 模型。第一个任务是训练模型并准备最终流程的模型。收集 SciTLDR 数据集用于训练模型。然后，准备将数据集用作所选模型的训练数据。添加特定于任务的前缀以总结各个论文。然后根据需求对模型进行微调。然后使用训练数据训练模型并预测结果。结果即为各个论文的摘要。然后使用 ROUGE 分数进行评估，并保存模型以供后续系统流程使用。Transformer 模型的训练概览如图 3 所示。

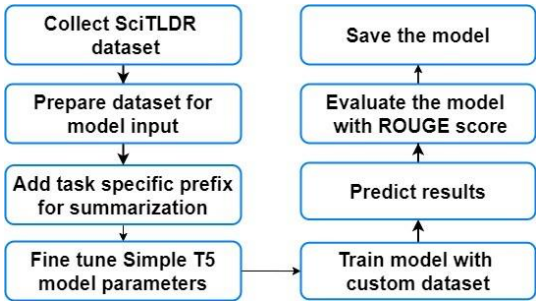

图3：**Transformer 模型的训练**

下一步是使用基于 Transformer 的模型实现系统流程，自动生成文献综述片段。该系统以多篇论文的 DOI 和 PDF 文件作为输入。它使用 Requests 库从 DOI 中收集论文标题和第一作者姓名。然后，它使用 PYPDF2 和正则表达式 (RE) 库收集每个 PDF 文件的摘要、引言和结论。之后，它将这三个部分合并起来，得到最终的模型输入。之后，它使用之前训练并保存的 T5 模型来获取每篇论文的摘要。下一步，它会进行后处理，并合并所有摘要，以生成连贯的文献综述片段。Transformer 模型的系统流程概览如图 4 所示。

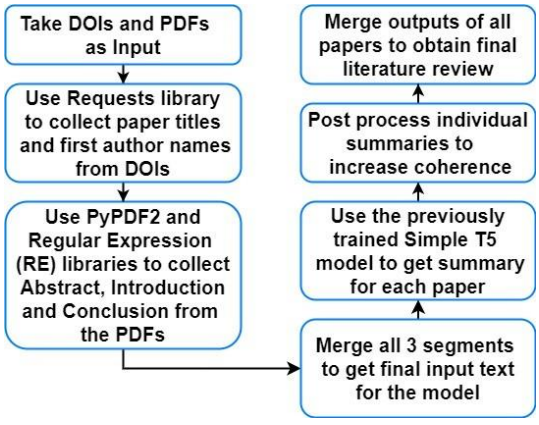

图4：**使用 Transformer 模型的管道**

D. 利用大型语言模型的程序：GPT-3.5-TURBO-0125

第三个流程采用基于 RAG 的方法，并使用大型语言模型：GPT-3.5-TURBO-0125。第一个任务是创建自定义 OpenAI 助手。首先，收集 SciTLDR 数据集，然后为 OpenAI 助手选择 GPT-3.5-TURBO-0125 模型。开启检索功能，并将数据集添加到 LLM 的知识库中。现在进行一些快速工程以生成所需的输出。然后，使用 ROUGE SCORE 评估 LLM 结果。OpenAI 助手的创建过程如图 5 所示。

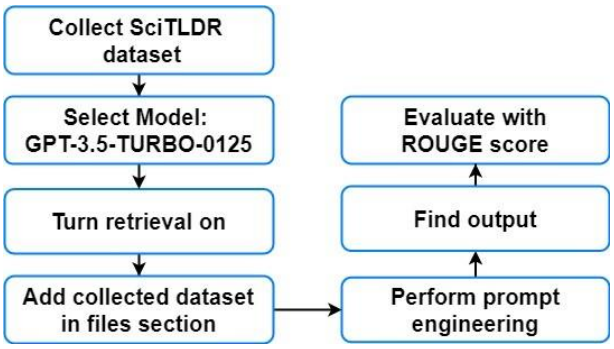

图5：**创建自定义 OpenAI 助手**

使用的提示：“用户将向您提供一个pdf文件作为输入，类似于知识库中给定的“data.json”文件的“输入”字段。您必须根据所提供文件为您的知识生成给定pdf的摘要“输出”。输出最多为80字。注意：您必须以一种可以被视为新研究论文的文献综述的方式撰写。用户将来可能会添加更多PDF，因此请尝试使文献综述连贯并符合IEEE标准。请提及第一作者的姓名和论文标题。不要像这样写“...的文献综述”。

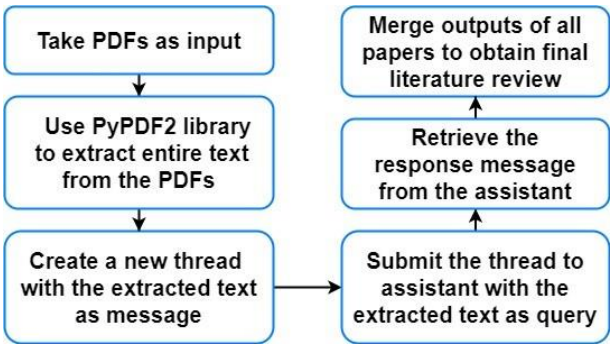

图6：**使用 LLM 的管道**

下一步是使用 LLM 自动生成文献综述片段，实现系统流程。该系统以多篇论文的 PDF 文件作为输入。它使用 PYPDF2 库提取每个 PDF 文件的全文。然后，它会创建一个新的线程，将提取的文本作为消息，并将提取的文本作为查询提交给助手。然后，检索助手的响应，并将每篇论文的输出合并，形成最终的文献综述部分。LLM 的系统流程概览如图 6 所示。

E. 最终系统工具

最终系统使用大型语言模型 GPT-3.5-TURBO-0125 作为后端实现。系统创建了一个美观简洁的用户界面，用户可以轻松上传多篇研究论文或 PDF 文件。用户只需点击“浏览文件”按钮，然后选择要上传的文件即可。之后，系统会加载研究论文，并在几秒钟内自动生成文献综述片段。系统会单独处理每篇论文并生成输出。加载界面和处理文件编号指示进度级别和已处理的论文数量。文献综述结束时，用户界面会显示“完成”文本，以指示任务完成。系统的用户界面如图 7 所示。

图7：**系统UI预览**

IV. S系统埃估值

本研究使用 ROUGE 评分进行评估。评估基于所选数据集的测试数据进行。ROUGE（面向召回率的摘要评估替代研究）是一组用于评估机器生成摘要质量的指标，通过将机器生成的摘要与参考摘要进行比较来评估其质量。使用的 ROUGE 指标包括：

- ROUGE-N（n-gram 重叠的准确率、召回率和 F1 分数），
- ROUGE-L（测量最长公共子序列）
- ROUGE-Lsum（ROUGE-Longest 用于摘要级别评估）

A.基于频率的 spaCy 评估

利用 ROUGE 分数在测试数据上对基于 spaCy 的模型进行了评估。结果如表 1 所示。

表一：**spaCy 的 ROUGE 分数**

ROUGE-1	0.257
ROUGE-2	0.055
ROUGE-L	0.144
红杉	0.146

B. 变压器评估

T5 利用 ROUGE 分数在测试数据上对基于 Transformer 的模型进行了评估。结果如表 II 所示。

表二：**T5 的 ROUGE 分数**

ROUGE-1	0.268
ROUGE-2	0.115
ROUGE-L	0.204
红杉	0.204

C.大型语言模型评估：GPT-3.5-TURBO-0125

利用ROUGE分数对基于LLM的模型在测试数据上进行了评估，结果如表3所示。

表三：**法学硕士 (LLM) 的 ROUGE 分数**

ROUGE-1	0.364
ROUGE-2	0.123
ROUGE-L	0.181
红杉	0.182

D. 多种方法的比较

表 IV 提供了这三种方法的比较。

表IV：**Rouge 评分比较**

	T5	spaCy	GPT-3.5-TURBO-0125
ROUGE-1	0.268	0.257	0.364
ROUGE-2	0.115	0.055	0.123
ROUGE-L	0.204	0.144	0.181
红杉	0.204	0.146	0.182

从 ROUGE 得分来看，基于 LLM 的模型明显优于 T5 和 spaCy。基于 Transformer 的模型在 ROUGE-1 得分中排名第二。

分数。评估证明，大型语言模型优于旧的NLP方法。

E. 用户对UI的评价

用户评估是在基于 LLM 的 UI 系统上进行的。用户选择了 4 篇研究文章作为输入。这些文章分别是：

- 基于AI-IoT的医疗预测交互系统
- 物联网和机器学习在个性化医疗中的应用：问题与挑战
- 利用物联网改善患者医疗保健
- 基于物联网的人工智能及其在工业中的应用

系统收到的输出：“Joshua Ernest Pedi Reddy 等人的论文提出了一种基于人工智能-物联网 (AI-IoT) 的医疗预后交互系统 (HPIS)，可在医护人员缺席的情况下提供实时医疗诊断和支持。该系统包含一个基于人工智能的聊天机器人和应用程序界面，用于收集信息、解答医疗疑问、协助用户并确保患者依从性。该系统专注于提高医疗服务的可及性和质量，集成了智能药丸分配器，用于监测和提高患者依从性，为医疗服务的革新提供巨大潜力。Ahamed 探索了物联网和机器学习在个性化医疗 (PH) 中的应用，以增强疾病管理和干预。物联网传感器设备和可穿戴设备收集患者数据，并使用人工智能和机器学习技术进行分析，以进行疾病预测和患者自我管理。挑战包括数据收集存在偏差、训练数据集过时以及隐私问题。物联网和机器学习在 PH 中的集成面临着数据传输可靠性和训练数据集存在偏差等问题，这些问题会影响诊断的准确性。应对这些挑战对于改进个性化医疗系统至关重要。” Khurana 实施了一套智能医疗系统，利用物联网传感器来提升医院的患者护理水平。该系统利用连接到 Arduino Uno 的超声波和红外接近传感器，实现静脉输液液位的自动监测和患者报警系统。文献综述强调了自动化医院管理系统对高效医疗管理的重要性。该系统旨在减少人力、成本和人为错误，同时改善患者护理。未来的研究旨在增强系统的可扩展性和功能性，以使其在医院得到广泛应用。Sherif El-Gendy 在论文《基于物联网的人工智能及其在工业中的应用》中探讨了物联网和人工智能在工业中的融合。该论文深入探讨了工业 4.0、工业物联网 (IIoT)、工业物联网 (IAIoT) 和物联网实时传输 (IoRT)，展示了它们对自动化和机器人技术的影响。论文探讨了物联网的挑战、人工智能在数据分析中的优势，并展示了 ABB 和波音等公司在油田生产优化和智能机器人技术方面的案例研究。物联网/人工智能融合的未来将为各个领域带来变革性进步。

V. R结果和D讨论

本研究介绍了三种自动生成文献综述的程序。研究还比较了各种自然语言处理方法之间的性能。

例如基于频率的方法 (spaCy)、Transformer 模型 (Simple T5) 以及基于 LLM (GPT-3.5-turbo) 的检索增强生成 (RAG)。这三种方法均已实现，并基于测试数据集计算了 ROUGE-1、ROUGE-2、ROUGE-L 和 ROUGE-Lsum 分数。对于这三种方法，ROUGE-1 和 ROUGE-2 分数均高于可接受水平。

从评估结果来看，GPT-3.5-turbo 模型的 ROUGE-1 和 ROUGE-2 得分均高于 SpaCy 和 T5。LLM 的总体 ROUGE-1 得分为 0.364，而 T5 的得分为 0.268，spaCy 的得分为 0.257。这表明 LLM 生成的摘要与人工摘要的单字和双字重叠度更高。Transformer T5 也是一款先进的模型，排名第二。排名最后的是基于频率的 spaCy 模型。

从得分来看，最先进的模型显然是 LLM，其表现优于所有其他 NLP 技术。但其他方法，例如 Transformer 模型和基于频率的方法，也能够产生令人满意的 ROUGE 得分和连贯的文献综述片段。

VI.C结论和F未来秒科佩斯

本研究重点实施并比较了各种用于自动文献综述的自然语言处理 (NLP) 技术。所有三个已实施的系统均成功生成了研究论文中连贯的文献综述部分。此外，还成功获取并比较了各种自然语言处理技术（例如基于频率的方法、Transformer 模型和大型语言模型）的结果。基于比较结果，基于 LLM 的方法被证明是 ROUGE-N 评分中表现最佳的方法。

因此，基于 LLM，还成功开发了一个最终的系统工具，用户可以上传多个 PDF 文件以自动生成连贯的文献综述片段。

本研究的未来工作可以集中于提高所开发系统工具的有效性和适用性。图形界面可以添加更多功能。

用户界面，例如模型选项、输出大小等。可以利用更多模型（例如 Bert、Gemini 和 LLaMA）来获得更好的结果。

R参考文献

[1] Felizardo KR，Carver JC。自动化系统性文献综述。当代软件工程中的实证方法。2020：327-55。

[2] Adhikari S. 基于自然语言处理（NLP）的机器学习文本摘要方法。2020 年第四届计算方法与通信国际会议（ICCMC）于2020年3月11日举行（第535-538页）。IEEE。

[3] Cachola I, Lo K, Cohan A, Weld DS. TLDR: 科学文献的极端概括. arXiv 预印本 arXiv:2004.15011. 2020 年 4 月 30 日。

[4] Jugran S, Kumar A, Tyagi BS, Anand V. 使用 Python 和 NLP 中的 SpaCy 进行自动文本摘要。2021 年先进计算与工程创新技术国际会议 (ICACITE) 将于 2021 年 3 月 4 日举行（第 582-585 页）。IEEE。

[5] Ali NF, Tanvin JU, Islam MR, Ahmed J, Akhtaruzzaman M. Google T5 与 SpaCy 在 YouTube 新闻视频摘要中的 ROUGE 评分分析及性能评估。2023 年第 26 届国际计算机与信息技术会议 (ICCIT)，2023 年 12 月 13 日（第 1-6 页）。IEEE。

[6] da Silva Ju ńior EM，Dutra ML。系统文献综述自动撰写路线图。《伊比利亚美洲科学测量与传播杂志》。2021年7月27日。

[7] Yuan W, Liu P, Neubig G. 我们能实现科学评审自动化吗？. 人工智能研究杂志。2022年9月29日;75:171-212。

[8] Karakan B、Wagner S、Bogner J. 系统文献综述的工具支持：分析现有解决方案和自动化潜力（斯图加特大学博士论文）。

[9] Jaspers S, De Troyer E, Aerts M. 机器学习技术在EFSA文献综述和系统评价自动化中的应用。EFSA支持出版物。2018年6月;15(6):1427E。

[10] Tauchert C、Bender M、Mesbah N、Buxmann P. 迈向使用机器学习进行自动化文献综述的综合方法。

[11] Tsai HC, Huang YF, Kuo CW. 基于 Mistral 大型语言模型与人工审阅的自动文献评审比较分析。

[12] Susnjak T、Hwang P、Reyes NH、Barczak AL、McIntosh TR、Ranathunga S. 通过特定领域大型语言模型微调实现研究综合自动化。arXiv 预印本 arXiv:2404.08680。2024 年 4 月 8 日。

[13] AllenAI。SCITL-DR 数据集。[数据集]。Hugging Face。[在线]。获取方式：https://huggingface.co/datasets/allenai/scitldr。[访问日期：2024 年 9 月 8 日]。