一、项目概述

1. 研究背景

随着大语言模型在自然语言处理领域的广泛应用和发展，其在问答系统和文本生成等一些任务中展现出了卓越的性能。然而由于模型的规模和复杂性，大语言模型生成的回答往往缺乏透明度和可验证性，很难获取它们产生响应的原因，这降低了终端用户对于大语言模型应用程序的信任和信心，在用于问答任务的检索增强生成（RAG）领域中也是如此。目前在使用 RAG 技术实现问答任务时，有的研究工作让 LLM 生成带有引文的文本，方便用户验证回复的内容，但是现有的归因方法普遍存在引文冗余和粒度粗糙等问题，很多情况下生成的引文并不足以得出最终答案，仍需用户仔细阅读其上下文。这导致引文的可理解性降低，致使用户在验证答案时需耗费额外精力筛选有效信息，进而影响用户对模型所生成回复的准确性和可靠性进行判断。

为了缓解以上问题，本文为使用 RAG 技术实现的大模型问答系统设计一种自动归因框架，该框架能生成准确且细粒度、更利于人类理解的引文。该框架在初期只需要少量的人工标注样本，可以让开源大模型自动生成带引文的样本来扩充数据集。在实现上，本文首先构建了一个人工标注的少量带细粒度引文的数据集，利用该数据集对模型进行微调，再用得到的模型对开源大模型生成的数据集进行打分，获取大量得分高的数据。最后利用该大规模数据微调了大模型 LlaMa3.1-8b 以及 Qwen2.5-7b，成功实现了在单个输出中生成精确的回复和更佳细粒度的引文。实验方面，在人工标注的带引文的 QA 问题上进行了评估，结果表明所训练的模型能够生成优质的引文。

2. 研究目标

本次结课大作业设计的目标是构建一个基于大模型的迭代优化的细粒度引文生成框架，提升 RAG 系统中答案归因的精确性和可读性。针对当前归因方法存在的引文冗余、可读性差等问题，研究把人类友好型标注原则贯穿整个流程，促使生成的引文准确又符合人类理解习惯。

3. 核心贡献

我们研究了基于 LLM 的 RAG 系统的简洁和充分的子句级引用的生成，重点是与传统阅读模式的一致性。我们提出了一套反映我们引用标准的注释原则，并相应地构建了一个手动注释的数据集。

我们设计了一种生成此类引用的方法，只需要少量手动注释的训练示例。

我们用我们构建的数据集进行实验，以验证我们方法的有效性。

二、数据来源与处理

1. 数据来源

在 RAG 归因以及 QA 问答任务的发展过程中，研究者们已经总结并整理出了多领域的数据集以供使用。本文选择四个经典的高质量数据集进行初始的人工标注归因实验，分别为：XQUAD、XOR-AttriQA、ELI5 以及 ASQA。接下来将详细介绍它们。

⑴ XQUAD

XQUAD 是跨语言问答数据集，包含 10 种语言的验证集样本，用于评估模型在多语言场景下结合上下文回答问题的能力，支持英语和德语任务验证。

⑵ XOR-AttriQA

XOR-AttriQA 聚焦开放域问答的答案归因，覆盖孟加拉语、芬兰语等五类语言，通过人工标注的答案-文档支持关系检验归因方法的跨语言鲁棒性。但是要注意的是原始归因都是句粒度级，本次实验是人工筛选其中可以精简，并且精简后效果更好的例子纳入初始的标注数据中。

⑶ ELI5

ELI5 数据集基于 Reddit 社区的″Explain Like I'm Five″板块构建，要求模型从 Sphere 语料库（过滤版 Common Crawl）中整合多文档生成解释性长答案，强调可验证引用支持。初始的数据中有部分是通过截取该数据集中的部分文档作为文章内容的。

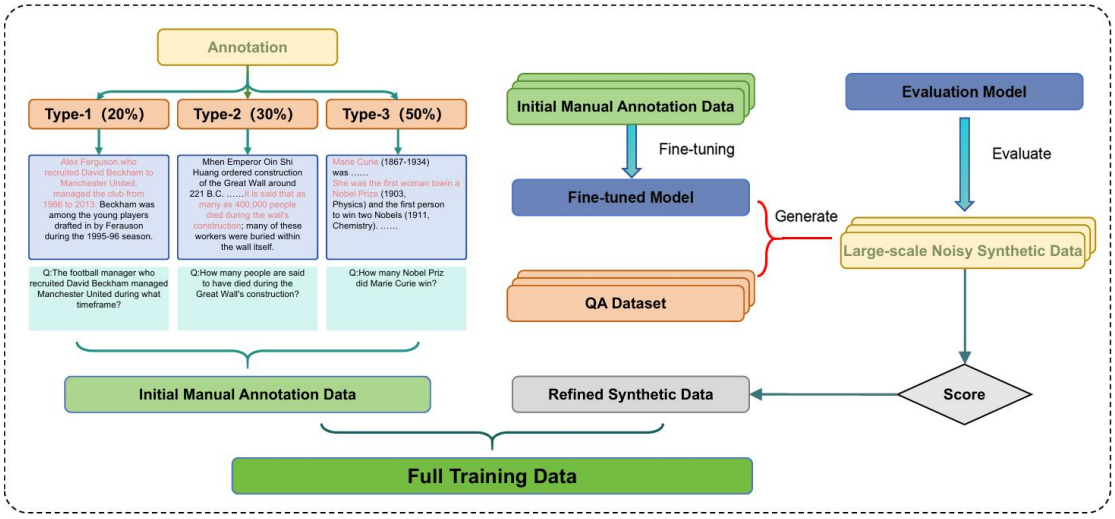
⑷ ASQA

专为处理模糊事实型问题设计，基于 AmbigQA 构建。每个问题对应多个潜在答案，要求模型生成覆盖多角度的长形式回答。数据源自 Wikipedia，通过人工整合分散答案形成标准回复，用于评估模型综合信息及生成全面解释的能力。其核心挑战在于消解歧义并确保答案完整性，常作为检索增强生成（RAG）系统的基准测试任务。

在数据集初期构建阶段，为了保证数据的质量保持在较高的水平，是通过人工从以上四个数据集中筛选可以利用的例子并且截取、标注好最后加入数据集

1.0 中作为初始的人工标注数据集以进行下一步的实验。每一条标注数据都包含四个字段，分别是文章原文、问题、答案、以及人工标注的细粒度引文。

三、方法



基于前章节中描述的注释原则，我们引入了一个数据增强框架，为我们的子句级引文生成任务自动构建训练实例。该框架可用于解决高注释成本导致的手动注释数据稀缺的问题。该过程如图所示。它由两个主要步骤组成：1）使用微调的自动注释模型生成实例候选；2）用评估模型过滤这些候选者。

首先，我们使用一个小规模的手动注释训练集对 LLM 进行监督微调（SFT），以开发一个自动注释模型。训练实例被构造为输入-输出对。输入包括一个问题、其相应的上下文，以及模型生成答案和提取支持事实的指令。输出由基本事实答案和子句级支持事实组成。然后将该模型应用于大量（问题、上下文）对，以生成一个以机器生成的答案和引用为特征的大规模数据集。由于该过程的自动化特性，产生的数据不可避免地包含噪声并表现出质量不一致。

为了提高生成的训练集的质量，我们实现了后续的过滤阶段。首先，我们丢弃机器生成的答案与地面真实参考不一致的实例。然后，使用二级 LLM 作为评估器来评估机器注释的实例。每个样本都根据一组标准进行评分，包括对源上下文的事实保真度、简洁性、答案验证的充分性和整体连贯性。仅保留超过预定义质量阈值的实例。

最后，手动注释的数据和保留的机器生成的实例共同构成了一个复合训练集。为了开发最终的引文生成模型，我们采用了一个两阶段的顺序 SFT 管道，首先对大规模合成数据进行微调，然后对高质量的手动数据进行细化。使用与自动注释模型相同的输入输出结构。

四、与现有方法对比

当下大语言模型也就是 LLMs 的可解释性以及可信度评估已然成为全球学术界与产业界都极为关注的热点问题，随着模型规模持续扩大以及应用场景不断拓展，怎样去理解模型内部的决策机制、评估生成内容的可靠性，以及建立有效的归因方法，这些共同构成了此领域研究的核心议题。近些年来，国内外学者于这些方向取得了一系列关键进展，推动了相关理论方法朝着创新发展以及实践应用的方向前进。

在模型可解释性研究领域，Reduan Achtibat 等人所提出的 AttnLRP 也即 Attention-Aware Layer-Wise Relevance Propagation 方法代表了当前最为前沿的技术进展。该方法创新性地对分层相关性传播也就是 LRP 技术进行了扩展，专门针对 Transformer 架构里的注意力机制设计了全新的归因规则，解决了传统方法在处理非线性注意力操作时所面临的挑战，AttnLRP 凭借深入剖析模型内部神经元活动，可解释输入特征的关键性，还可以揭示潜在表示层的决策依据，为理解 LLM 的复杂推理过程提供了新的视角。实验证明，该方法在 LlaMa 2、Mixtral 8x7b 等主流模型上的解释效果要明显优于传统技术，同时在计算效率方面表现优异，仅仅需要相当于一次反向传播的计算开销便可完成全面分析，这些优势让其成为当前极具实用价值的模型解释工具之一。

检索增强生成即 RAG 系统的可信度评估研究也有了关键进展，Yujia Zhou 团队构建的六维评估框架，囊括事实性、鲁棒性、公平性、透明度、问责制和隐私等关键指标，为 RAG 系统的评估提供了方法论支撑，该团队开发的评估基准，对专有和开源模型等多种系统进行了横向比较，指出当前技术在实际应用中的局限。研究发现，虽然 RAG 机制凭借引入外部知识缓解了模型幻觉问题，但检索信息的质量控制和合理利用仍是挑战，在医疗、法律等高风险应用场景中，不恰当引用可能导致严重后果，这突出了建立严格可信度标准的关键性。

在答案归因技术领域当中 SelfCite 方法的出现意味着自监督学习在这个应用里成功实践，Yung-Sung、Chuang 等人设计的上下文消融奖励机制巧妙运用模型自身能力来进行质量评估，并不依赖昂贵的人工标注，借助最佳 N 采样和偏好优化相结合的策略，此方法在 LongBench-Cite 基准上把引用质量提升了 5.3 个 F1 分数点，显示出自监督学习在提升模型可解释性方面有很大潜力。Viju Sudhi 团

队开发的 RAG-Ex 框架提供了一种不依赖模型的解释方案，其基于扰动测试的特征关键性分析方法能兼容各类 LLM 架构，在多种语言任务里保持稳定的解释性能，用户研究说明该框架生成的解释与人工判断的一致性达到 76.9%，给终端用户理解模型行为提供了可靠工具。Jonas Wallat 等人对引用正确性与忠实性的区分研究加深了对归因质量的理解，他们凭借严谨的实验设计指出“后合理化”现象普遍存在，就是模型大多时候会生成看似合理但实际没有忠实反映决策过程的引用，这种表里不一的归因行为可能让用户产生错误信任，在关键应用中造成严重风险。研究发现当前系统中高达 57%的引用存在忠实性问题，这个结果警示研究社区要建立更严格的评估标准，同时推动开发真正能反映模型决策过程的归因方法，Jirui Qi 团队提出的 MIRAGE 方法代表了基于模型内部信息的归因技术的最新发展，该方法借助分析上下文敏感标记与检索文档的关联关系，达成了细粒度的答案支持证据定位。在多语言抽取式 QA 任务中 MIRAGE 呈现出与人类判断高度一致的解释能力，在开放式生成场景下其性能与自引用方法差不多，同时提供了更灵活的参数调控空间，这些优势使它成为平衡解释质量和计算效率的理想选择。

国内学者于长上下文生成任务的归因技术领域有着关键贡献，Zhang Jiajie 团队所开发的 LongCite 系统成功达成了在超长文本里准确定位支持证据的技术进展，其构建的 CoF 流程创新性地运用从粗粒度至细粒度的渐进式归因策略，借由大规模监督微调让模型掌握了精确的句子级引用能力。基于 44,600 个高质量训练实例构建而成的 LongCite-45k 数据集，为后续的研究提供了珍贵资源，实验显示，经过专门训练的 LongCite-8B/9B 模型在引用质量方面超过了 GPT-4o 等商业系统，同时维持了更为精细的引用粒度，呈现出专业化训练在提升模型可解释性方面的优势。

当前的研究显现出一些需要解决的关键问题，在技术层面，现有的解释方法面对超大规模模型时，在计算效率上仍存在挑战，难以实现实时且全面的分析，多数方法只能解释模型的局部行为，缺乏对复杂推理链的完整追踪能力，在评估标准方面，虽然已经出现了如 LongBench-Cite 等基准测试，但是社区尚未构建统一的评估协议，不同研究之间的结果可比性不足。针对中文等非拉丁语系语言的专门研究相对匮乏，现有方法在跨语言迁移中的适应性有待检验。

五、实验结果

1. 实验设置

我们使用初始学习率为 5e-5 的 AdamW 优化器和余弦学习率调度器，用梯度裁剪（最大范数 1.0）训练 4 个迭代周期。为了提高计算效率，我们利用 bf16 混合精度训练。实验是在每个 GPU 的批处理大小为 2 和 8 个梯度累积步骤的情况下进行的。进行了参数高效微调，对于 LoRA 配置，我们设置了秩 r=8 和缩放因子 alpha=32，且零丢包。此外，我们采用 LoRA+ 优化器设置，B 矩阵学习率比为 16。

Model	Strategy	Performance Metrics (%)					
		R	P	F1	ROUGE-L	BLEU	CL
LlaMa3.1-8b	Base	26.85	18.09	20.61	58.49	43.87	142.21
	Synthetic Only (10K)	77.49	78.84	77.53	**87.32**	**79.64**	82.44
	Manual Only (0.5K)	75.81	76.64	75.42	86.37	77.91	83.52
	Full Method (10K+0.5K)	78.59	78.77	77.96	87.25	78.81	79.89
	Manual Only (1K)	77.22	77.76	76.92	86.87	78.60	82.01
	Full Method (10K+1K)	**79.61**	**78.93**	**78.61**	87.27	78.98	81.42
Qwen2.5-7b	Base	27.57	28.71	27.89	63.83	50.09	137.07
	Synthetic Only (10K)	75.88	76.92	75.74	86.15	77.97	81.01
	Manual Only (0.5K)	74.06	74.59	73.00	83.60	73.76	83.14
	Full Method (10K+0.5K)	76.39	76.08	75.50	86.15	78.00	83.16
	Manual Only (1K)	74.81	75.27	74.21	84.87	75.64	82.86
	Full Method (10K+1K)	**77.78**	**77.61**	**76.99**	**86.58**	**78.12**	78.93
GLM-4-9b	Base	34.47	35.77	34.74	59.63	47.81	128.32
	Synthetic Only (10K)	74.65	**75.62**	74.47	83.48	74.66	83.52
	Manual Only (0.5K)	69.42	69.33	68.63	82.61	73.39	94.36
	Full Method (10K+0.5K)	74.46	74.65	73.92	83.73	75.54	86.69
	Manual Only (1K)	74.87	74.63	74.11	84.89	76.44	89.47
	Full Method (10K+1K)	**75.79**	74.82	**74.63**	**85.68**	**77.34**	86.09

所有实验均在监督微调（SFT）范式下在多个基础模型（Llama-3.1-8B、Qwen2.5-7B、GLM-4-9B）上进行，采用低秩自适应（LoRA）进行轻量级自适应。关键实现细节和超参数配置如下。

我们验证了我们的数据增强框架和两阶段微调范式的有效性。结果如表所示。

在表中，Base 表示没有任何微调的原始 LLM 模型，它仅通过提示工程以目标格式生成输出。Manual Only 表示仅在人类注释实例上微调的模型。Synthetic Only 是仅使用自动注释进行微调的模型。Full Method 表示使用我们的两阶段策略训练的模型，该策略将 10000 个机器生成的实例与相同的人类注释样本结合在一起。

原始基础模型在子句引用生成方面表现出明显的不足，而所有微调的变体都取得了显著的性能提升，两阶段微调策略产生了最佳结果。10K+1K 两级方案在几乎所有指标和实验配置中都能始终如一地产生优异的结果。

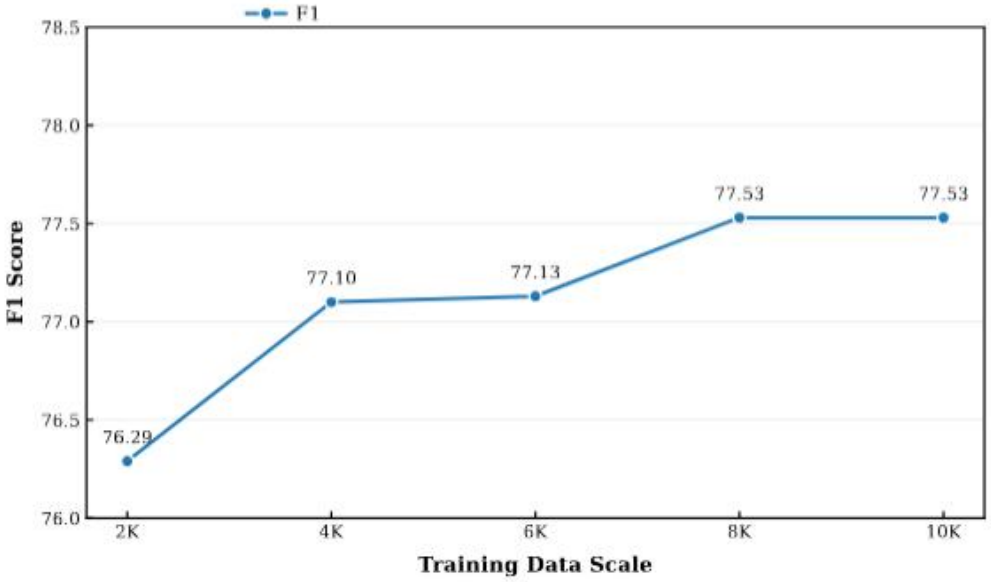
值得注意的是，使用 10K 机器生成数据的单级微调优于仅依赖人类注释数据（0.5K 或

1K）的基线。这一结果突显了通过我们的框架构建的机器生成数据的扩展优势。

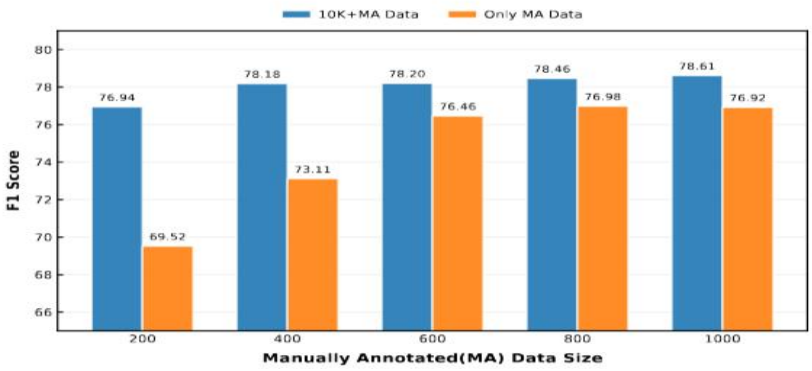
对基本模型和微调模型之间的引用长度（CL）进行比较后发现，微调后引用长度显著减少。这种转变表明，模型学会了生成更简洁的引用。

不同基础模型之间的性能差异突显了我们的微调策略如何适应不同的架构特征。这些结果支持了两个关键的观察结果。首先，原始基础模型通常难以生成高质量的子句引用，因为它们缺乏针对这种细粒度归因任务的特定任务优化。其次，我们的两阶段微调策略有效地利用了机器生成数据的规模优势和人工注释数据的质量优势，从而显著改善了核心评估指标。

2. 消融实验



我们分析了训练数据大小对模型性能的影响，重点关注两种情况：使用大规模机器生成数据的单阶段微调和辅以人工注释数据的两阶段微调。图显示了随着机器生成的数据量从 2K 增加到 10K，模型性能的缩放行为。我们观察到 F1 得分上升到 8K。然而，超过这个阈值后，性能增长开始趋于平稳。这表明 8K-10K 范围是数据生成的一个经济高效的规模。



上图说明了在二次微调阶段用 200-1000 个人工注释实例补充 10K 机器生成的基础对性能的影响。结果表明，对合成数据进行预训练可以持续提高最终表现。值得注意的是，F1 分数的提高在低资源场景（例如 200 个人类实例）中最为明显，并且随着人类注释数据量的增加而逐渐减少。

Method	F1
Longcite-llama3.1-8b	45.29
ALCE	52.46
Rag-ex 2.0	58.91
Ours	**78.61**

如上表所示，所提出的框架在很大程度上超越了现有的方法，证明了其在子句级引文生成方面的优越性。基线方法的性能较低是意料之中的，因为它们没有针对这项任务的细微差别进行专门优化。

Methods	Q1: "What is the capital of the state of Assam?" Answer: "Dispur."	Q2: "On what OS was the YouTube Kids app available?" Answer: "Android and iOS"
	Generated Citations	Generated Citations
RAG-Ex 2.0	Dispur is the capital of the state of Assam in India. Dispur, a locality of Guwahati, became the capital of Assam in 1973. This was after Shillong, the erstwhile capital,became the capital of the state of Meghalava that was carved out of Assam. Dispur is the seat of Government of Assam. …	In February 2015, YouTube announced the launch of a new app specifically for use by children visiting the site, called YouTube Kids. It allows parental controls and restrictions on who can upload content, and is available for both Android and iOS devices. Later on August 26, 2015, YouTube Gaming was launched…
Longcite-LlaMa3.1-8b	Dispur is the capital of the state of Assam in India. Dispur, a locality of Guwahati, became the capital of Assam in 1973. This was after Shillong, the erstwhile capital,became the capital of the state of Meghalava that was carved out of Assam. Dispur is the seat of Government of Assam. …	In February 2015, YouTube announced the launch of a new app specifically for use by children visiting the site, called YouTube Kids. It allows parental controls and restrictions on who can upload content, and is available for both Android and iOS devices. Later on August 26, 2015, YouTube Gaming was launched …
ALCE	Dispur is the capital of the state of Assam in India. Dispur, a locality of Guwahati, became the capital of Assam in 1973. This was after Shillong, the erstwhile capital,became the capital of the state of Meghalava that was carved out of Assam. Dispur is the seat of Government of Assam. …	In February 2015, YouTube announced the launch of a new app specifically for use by children visiting the site, called YouTube Kids. It allows parental controls and restrictions on who can upload content, and is available for both Android and iOS devices. Later on August 26, 2015, YouTube Gaming was launched
Ours	*Dispur is the capital of the state of Assam in India.* Dispur, a locality of Guwahati, became the capital of Assam in 1973. This was after Shillong, the erstwhile capital,became the capital of the state of Meghalava that was carved out of Assam. Dispur is the seat of Government of Assam. …	*In February 2015, YouTube announced the launch of a new app specifically for use by children visiting the site, called YouTube Kids. It allows parental controls and restrictions on who can upload content, and is available for both Android and iOS devices.*Later on August 26, 2015, YouTube Gaming was launched…

如表所示，展示了两个具有代表性的开放域 QA 场景中不同方法的引用输出。

Q1 涉及阿萨姆邦首府，涉及一个包含多个句子的上下文，可以独立支持答案。第二季度，关于 YouTube Kids 支持的操作系统，存在高密度的无关背景信息。这两个场景测试了方法在支持事实提取时平衡信息性和简洁性的能力。对于第一季度，RAG Ex 2.0 的单句扰动策略未能产生任何支持事实。这是因为当多个句子可以独立地支持答案时，删除任何一个句子仍然会留下足够的信息，导致可以忽略的扰动差异，从而无法识别核心支持内容。同时，监督方法 Longcite-LlaMa3.1-8b 和 ALCE 表现出显著的引用冗余。他们的输出包括多个支持句，即使一个引用就足够了，从而增加了用户的认知负荷，他们必须在重复的声明中验证答案。对于 Q2，上下文包含与操作系统支持问题无关的大量无关细节。RAG Ex 2.0、Longcite-LlaMa3.1-8b 和 ALCE 都未能过滤这些干扰：它们的引用保留了非核心内容以及关键信息，使支持事实难以快速定位。

相比之下，我们的方法根据每种情况的需求量身定制引用：对于第一季度，我们只保留了一个核心声明，确认 Dispur 是阿萨姆邦的首都，不包括冗余的重复和无关的背景。对于第二季度，我们去掉了非必要的应用程序功能细节，只关注操作系统支持信息。这确保了我们的引用既足以验证答案，又足够简洁，以避免信息过载。

这两种不同场景的并排比较证明了我们的子句级引文生成方法的优越性。它确定了验证所需的确切子句片段，在不牺牲信息量的情况下成功地消除了引文冗余。

六、总结与不足

在这项工作中，研究了基于 RAG 的系统的简洁而充分的子句级引用，以解决现有引用策略可能经常包含大量无关信息或遗漏一些验证所需内容的问题。我们制定了注释指南，并为改进的引用要求构建了一个数据集。提出了一种合成数据生成管道，该管道利用 LLM 生成具有严格自动过滤阶段的训练实例，以增强手动注释的数据。在我们构建的数据集上进行的实验验证了所提出方法的有效性。

我们的数据集缺少长形式实例。因此，直接应用所提出的方法为冗长的 LLM 回复生成引用，而不进行进一步的调整，可能会导致性能不佳。使用 LLM 生成引用内容会增加总运行时间。最后，由于引用的内容是由模型直接生成的，因此存在产生幻觉的风险，生成的引用可能与检索到的源上下文不完全一致。