一、项目概述

1. 研究背景

在检索增强生成（RAG）问答系统中，大型语言模型（LLM）生成的输出常存在幻觉问题，而引用机制是提升输出可验证性、帮助用户识别幻觉的关键手段。现有归因方法生成的引用多集中在句子或段落级别，存在两大核心缺陷：一是包含大量无关内容，增加用户验证成本；二是可能遗漏验证所需的关键信息（如指代对象），迫使用户查阅上下文。因此，亟需一种兼顾简洁性与充分性的引用生成方案。

2. 研究目标

设计并实现一种子句级引用生成方法，确保生成的引用内容自然连贯、简洁精炼，同时包含验证答案所需的全部关键信息，降低用户验证难度，提升 RAG 系统的可靠性。

3. 核心贡献

1.提出一套符合常规阅读习惯的子句级引用标注准则，构建了涵盖单句、去冗余、多片段引用场景的人工标注数据集。

2.设计了“少量人工标注 + 大规模数据增强”的框架，仅需 2K 人工标注样本即可生成高质量训练数据，解决数据稀缺问题。

3.提出两阶段微调策略，有效融合机器生成数据的规模优势与人工标注数据的质量优势，显著提升模型的子句级引用生成能力。

二、数据来源与处理

1. 数据来源

本实验采用三个互补的公开问答数据集作为原始语料库，覆盖不同 discourse 结构、内容多样性和推理深度需求：

1. XOR-AttriQA：提供开放域 QA 的归因标注，以句子级引用为主，经二次精炼后可获取与答案直接对应的细粒度片段；
2. XQUAD：段落级阅读理解基准数据集，补充常规问答场景的语料多样性；
3. HotpotQA：以多跳推理为核心特色，验证模型在长距离依赖和指代消解场景下的引用定位能力。

2. 数据处理流程

（1）标注原则与基准构建

基于 “自然连贯、简洁充分” 的核心要求，定义三类引用实例：

1. Type-1：引用内容与传统句子级标准一致，单句即可完整支撑答案验证；
2. Type-2：传统句子级引用包含无关内容，需裁剪冗余片段（常见于正式文档类语料）；
3. Type-3：需引用多个分散片段，包括“答案支持句未明确指代对象”和“答案需多跳推理”两种场景。人工标注 2000 个实例作为种子数据集，每个实例包含 “问题 - 答案 - 支持事实（子句级引用）- 类型标签” 四要素，确保标注质量符合基准要求。

（2）数据增强框架

为解决人工标注成本高、规模小的问题，设计三阶段数据增强流程：

1. 种子数据构建：手动标注 2K 高质量实例，作为模型学习的基准；
2. 大规模噪声数据生成：使用种子数据对开源 LLM 进行少样本微调，得到基线模型，再用该模型处理原始语料库中的问答对，生成大规模机器标注引用数据；
3. 质量过滤：采用 GPT-4 作为评估模型，从事实一致性、简洁性、验证充分性、连贯性四个维度对机器生成数据评分，筛选超阈值的高质量数据，最终形成大规模训练数据集（Ours New Dataset）。

三、方法

1. 核心框架

本实验的归因框架核心是 “数据增强 + 两阶段微调”，整体流程如下：原始语料库 → 人工标注 2K 种子数据 → LLM 生成大规模噪声数据 → GPT-4 过滤高质量数据 → 两阶段微调模型 → 子句级引用生成

2. 模型选择

实验选用三款主流开源 LLM 作为基础模型，验证方法的通用性：

1. LlaMa3.1-8b；
2. Qwen2.5-7b；
3. GLM-4-9b。

3. 微调策略

设计两类微调策略，对比不同数据使用方式的效果：

1. 单阶段微调：包含两种具体方案，分别是仅使用 0.5K/1K 人工标注数据、仅使用 10K 机器生成数据，核心逻辑是单独验证人工标注数据的质量优势与机器生成数据的规模优势；
2. 两阶段微调：具体方案是先以 10K 机器生成数据预训练，再用 0.5K/1K 人工标注数据二次微调，核心逻辑是结合规模优势与质量优势，先学习通用模式，再优化细粒度性能。

4. 评估指标

采用多维度评估体系，全面量化引用质量：

1. 词汇匹配指标：精确率（P）、召回率（R）、F1 分数（harmonic mean of P 和 R）；
2. 语义一致性指标：余弦相似度（CS）；
3. 序列匹配指标：ROUGE-L（最长公共子序列）、BLEU（n-gram 重叠）；
4. 主观质量指标：GPT-4o 评分（基于准确性、简洁性、可读性）；
5. 辅助指标：CL 指标（评估引用内容的合理性与冗余度）。

F1 分数的计算过程如下：首先，用正则表达式将文本拆分为词元序列，统一转为小写并消除空格差异，构建可比单元；随后，将预测序列和参考序列的词元分别构建为无序集合，消除词频干扰，仅保留形态信息；基于这些集合，计算两者的交集大小作为真阳性（TP）；精确率（P）定义为预测文本中正确词元的比例，即 TP 与预测集合基数的比值；召回率（R）定义为参考文本中被成功覆盖的词元比例，即 TP 与参考集合基数的比值；最终 F1 分数为精确率和召回率的调和平均数，公式为 F1=（2×P×R）/（P+R），综合反映词汇匹配的全面性和准确性。

四、实验结果

1. 模型性能对比

LlaMa3.1-8b 模型在不同策略下的性能指标如下：

- 基准模型（Base）：召回率（R）26.85%、精确率（P）18.09%、F1 分数 20.61%、余弦相似度（CS）0.6887、ROUGE-L 58.49%、BLEU 43.87%、GPT-4o 评分 82.53%、CL 指标 142.21；
- 0.5K 人工标注数据微调：召回率（R）75.81%、精确率（P）76.64%、F1 分数 75.42%、余弦相似度（CS）0.8835、ROUGE-L 86.37%、BLEU 77.91%、GPT-4o 评分 92.51%、CL 指标 83.52；
- 1K 人工标注数据微调：召回率（R）77.22%、精确率（P）77.76%、F1 分数 76.92%、余弦相似度（CS）0.8889、ROUGE-L 86.87%、BLEU 78.60%、GPT-4o 评分 93.23%、CL 指标 82.01；
- 10K 机器生成数据微调：召回率（R）77.49%、精确率（P）78.84%、F1 分数 77.53%、余弦相似度（CS）0.8915、ROUGE-L 87.32%、BLEU 79.64%、GPT-4o 评分 92.93%、CL 指标 82.44；
- 10K 机器生成数据 + 0.5K 人工标注数据两阶段微调：召回率（R）78.59%、精确率（P）78.77%、F1 分数 77.96%、余弦相似度（CS）0.8907、ROUGE-L 87.25%、BLEU 78.81%、GPT-4o 评分 93.00%、CL 指标 79.89；
- 10K 机器生成数据 + 1K 人工标注数据两阶段微调：召回率（R）79.61%、精确率（P）78.93%、F1 分数 78.61%、余弦相似度（CS）0.8928、ROUGE-L 87.27%、BLEU 78.98%、GPT-4o 评分 93.15%、CL 指标 81.42。

Qwen2.5-7b 模型在不同策略下的性能指标如下：

- 基准模型（Base）：召回率（R）27.57%、精确率（P）28.71%、F1 分数 27.89%、余弦相似度（CS）0.6831、ROUGE-L 63.83%、BLEU 50.09%、GPT-4o 评分 80.25%、CL 指标 137.07；

- 0.5K 人工标注数据微调：召回率（R）74.06%、精确率（P）74.59%、F1 分数 73.00%、余弦相似度（CS）0.8634、ROUGE-L 83.60%、BLEU 73.76%、GPT-4o 评分 92.53%、CL 指标 83.14；
- 1K 人工标注数据微调：召回率（R）74.81%、精确率（P）75.27%、F1 分数 74.21%、余弦相似度（CS）0.8728、ROUGE-L 84.87%、BLEU 75.64%、GPT-4o 评分 92.81%、CL 指标 82.86；
- 10K 机器生成数据微调：召回率（R）75.88%、精确率（P）76.92%、F1 分数 75.74%、余弦相似度（CS）0.8824、ROUGE-L 86.15%、BLEU 77.97%、GPT-4o 评分 92.61%、CL 指标 81.01；
- 10K 机器生成数据 + 0.5K 人工标注数据两阶段微调：召回率（R）76.39%、精确率（P）76.08%、F1 分数 75.50%、余弦相似度（CS）0.8831、ROUGE-L 86.15%、BLEU 78.00%、GPT-4o 评分 92.77%、CL 指标 83.16；
- 10K 机器生成数据 + 1K 人工标注数据两阶段微调：召回率（R）77.78%、精确率（P）77.61%、F1 分数 76.99%、余弦相似度（CS）0.8859、ROUGE-L 86.58%、BLEU 78.12%、GPT-4o 评分 92.63%、CL 指标 78.93。

GLM-4-9b 模型在不同策略下的性能指标如下：

- 基准模型（Base）：召回率（R）34.47%、精确率（P）35.77%、F1 分数 34.74%、余弦相似度（CS）0.6937、ROUGE-L 59.63%、BLEU 47.81%、GPT-4o 评分 78.35%、CL 指标 128.32；
- 0.5K 人工标注数据微调：召回率（R）69.42%、精确率（P）69.33%、F1 分数 68.63%、余弦相似度（CS）0.8499、ROUGE-L 82.61%、BLEU 73.39%、GPT-4o 评分 92.53%、CL 指标 94.36；
- 1K 人工标注数据微调：召回率（R）74.87%、精确率（P）74.63%、F1 分数 74.11%、余弦相似度（CS）0.8667、ROUGE-L 84.89%、BLEU 76.44%、GPT-4o 评分 92.40%、CL 指标 89.47；
- 10K 机器生成数据微调：召回率（R）74.65%、精确率（P）75.62%、F1 分数 74.47%、余弦相似度（CS）0.8394、ROUGE-L 83.48%、BLEU 74.66%、GPT-4o 评分 90.70%、CL 指标 83.52；
- 10K 机器生成数据 + 0.5K 人工标注数据两阶段微调：召回率（R）74.46%、精确率（P）74.65%、F1 分数 73.92%、余弦相似度（CS）0.8582、ROUGE-L 83.73%、BLEU 75.54%、GPT-4o 评分 89.15%、CL 指标 86.69；
- 10K 机器生成数据 + 1K 人工标注数据两阶段微调：召回率（R）75.79%、精确率（P）74.82%、F1 分数 74.63%、余弦相似度（CS）0.8687、ROUGE-L 85.68%、BLEU 77.34%、GPT-4o 评分 92.12%、CL 指标 86.09。

2. 训练数据量与性能趋势（曲线说明）

（1）训练数据规模对性能的影响

横轴为训练数据规模（2K、10K），纵轴分别为 F1 分数（左）和余弦相似度（CS，右）。随着训练数据规模从 2K 扩大到 10K，三款模型的 F1 分数和余弦相似度均呈上升趋势，验证了大规模数据对模型性能的提升作用；其中 LlaMa3.1-8b 的提升幅度最显著，说明其对数据规模的敏感度更高。

（2）两阶段微调数据量对性能的影响

横轴为第二阶段微调数据量（400、600、800、1000），纵轴分别为 F1 分数（左）和余弦相似度（CS，右）。在 10K 机器生成数据预训练的基础上，随着第二阶段人工标注数据量的增加，模型性能持续提升；当数据量达到 1000 时，性能趋于稳定，说明 1K 人工标注数据已能充分优化模型的细粒度引用生成能力。

3. 核心结论

1. 所有微调模型的性能均显著优于原始基线模型，其中 LlaMa3.1-8b 的 10K+1K 两阶段策略表现最优：F1 分数达 78.61%（较基线提升 3.8 倍），余弦相似度 0.8928（提升 29.6%），GPT-4o 评分 93.15%（提升 12.9%）。
2. 两阶段微调策略优于单阶段策略，验证了“规模数据预训练 + 质量数据微调”的合理性，既能利用机器生成数据的规模优势，又能通过人工标注数据修正偏差。

3. 单阶段 10K 机器生成数据的性能优于 0.5K/1K 人工标注数据，说明高质量机器生成数据的规模优势可部分弥补人工标注的质量优势。

五、问题分析与创新点

1. 现有方法的核心问题

1. 引用粒度粗：句子 / 段落级引用包含大量无关内容，增加用户验证时间；
2. 信息完整性不足：句子级引用可能遗漏指代对象等关键验证信息，需用户查阅上下文；
3. 解释性差：扰动或梯度 -based 归因方法生成的高亮元素分散，难以被用户高效理解；
4. 数据效率低：传统方法依赖大规模人工标注，成本高且难以扩展。

2. 本项目创新点

（1）引用粒度创新：子句级标注准则与数据集

首次明确子句级引用的 “自然连贯、简洁充分” 标注原则，覆盖单句、去冗余、多片段三类核心场景，构建了首个适配该需求的人工标注数据集，填补了细粒度引用数据集的空白。

（2）数据增强创新：少量人工 + 大规模生成的高效框架

设计 “人工标注种子数据→LLM 生成大规模数据→GPT-4 质量过滤” 的闭环流程，仅需 2K 人工标注样本即可生成高质量大规模训练数据，解决了细粒度任务的标注成本问题。

（3）训练策略创新：两阶段微调兼顾规模与质量

提出两阶段微调策略，第一阶段利用机器生成数据学习通用引用模式，第二阶段通过人工标注数据优化细粒度性能，有效平衡了数据规模与质量，提升了模型的泛化能力和准确性。

六、Demo 截图 / 链接

1. Demo 功能说明

Demo 实现 “输入问题 + 检索上下文→输出答案 + 子句级引用” 的端到端功能，核心展示引用的简洁性与充分性，支持高亮显示引用片段及其在原始上下文中的位置。

2. 示例演示

输入：问题为 “When was the Great Barrier Reef declared a UNESCO site?”，上下文为 “The Great Barrier Reef, located off the coast of Queensland, Australia, is the world's largest coral reef system. Composed of over 2,900 individual reefs and 900 islands spanning 2,300 kilometers, it supports extraordinary biodiversity including 1,500 fish species and 400 types of coral. Designated a UNESCO World Heritage Site in 1981, the reef faces threats from climate change, coral bleaching, and pollution.”

输出：答案为 “1981.”，子句级引用为 “The Great Barrier Reef; Designated a UNESCO World Heritage Site in 1981.”

3. 访问方式

当前 Demo 处于内部测试阶段，可联系项目团队获取测试链接，后续将开源部署至 GitHub。

七、未来改进方向

1. 跨模态归因支持：现有框架仅处理文本数据，未来将探索基于几何深度学习的跨模态对齐方法，将图像、音频等多模态信息映射到统一空间，实现多模态场景下的细粒度归因分析。
2. 超长文本处理优化：针对当前框架在超长文档中检索与引用生成效率不足的问题，将研究分段优化策略与注意力机制改进，提升模型在超长文本场景下的性能稳定性。
3. 多语言基准构建：现有评估指标主要适配英文 QA 场景，未来将建立包含多语言、多文本形式的归因基准测试集，提升方法的跨语言适应性。
4. 动态验证协议开发：基于因果推理设计动态验证协议，实时评估引用与答案之间的因果关联性，进一步提升评估体系的全面性与准确性。