摘要

检索增强型语言模型（RALMs）代表了大型语言模型能力的显著进步，特别是通过利用外部知识源来减少事实性错误。然而，检索到的信息的可靠性并不总是有保证的。不相关数据的检索可能导致误导性回应，并可能导致模型忽视其固有知识，即使在它拥有足够信息解决查询时也是如此。此外，标准RALMs通常难以评估它们是否拥有足够的知识（包括固有的和检索到的）来提供准确的答案。在缺乏知识的情况下，这些系统理想情况下应当在无法得到答案时回应“未知”。为了应对这些挑战，我们引入了一种名为“记录链”（CHAIN-OF-NOTING，简称CON）的新方法，旨在提高RALMs在面对噪声、不相关文档以及处理未知场景时的鲁棒性。CON的核心思想是为检索到的文档生成顺序阅读笔记，使得能够彻底评估它们与给定问题的相关性，并将这些信息整合以构建最终答案。我们利用ChatGPT来创建CON的训练数据，随后在LLaMa-2 7B模型上进行了训练。我们在四个开放领域的问答基准测试中的实验表明，配备了CON的RALMs显著优于标准RALMs。值得注意的是，CON在完全噪声的检索文档中平均提高了+7.9的EM（精确匹配）分数，并且在预训练知识范围之外的实时问题的拒绝率提高了+10.5。

引言

研究现状：检索增强型语言模型（RALMs）代表了一种新颖的框架，通过解决一些关键限制，如减少事实性幻觉、以即插即用的方式注入最新知识、增强领域专业知识，从而显著推进了大型语言模型的发展。这些增强主要来自于将大型语言模型与外部知识源整合。在典型的RALM设置中，首先通过检索器处理查询，检索器在庞大的证据库中搜索相关文档。然后，阅读器检查这些文档，提取有用信息并制定最终输出答案。RALM框架的潜在优势在于其能够整合相关的外部知识，从而丰富LLMs对输入文本的理解，并基于这些信息生成答案。当LLMs缺乏某一主题的直接知识时，这特别有益，允许它们以即插即用的方式获取并利用相关信息。

存在问题：（1）无法保证信息检索系统始终能够提供最相关或最可信的信息。检索不相关的数据可能导致误导性回应，并可能导致模型忽视其固有知识，即使在它拥有足够信息解决查询时也是如此。（2）处理面向事实问题时出现幻觉，这种缺陷可能是危险的，并可能会让用户感到气馁。理想情况下，一个智能系统应该能够判断它是否拥有足够的知识（包括固有的和检索到的）来提供准确的答案。在知识不足的情况下，如果无法确定答案，系统应该回应“未知”。

本文作者贡献：基于标准RALM系统的缺点，作者引入了一个名为CHAIN-OF-NOTING（CON）的新框架提高RALMs的鲁棒性，主要关注两个关键方面：（1）噪声鲁棒性：RALM识别并忽略检索到的不相关文档中存在的噪声信息的能力，同时适当利用其固有知识。（2）未知鲁棒性：RALM在给定一个它没有相应知识回答的查询时，通过回应“未知”承认其局限性的能力，以及在检索文档中未找到相关信息的情况下。作者对与CON集成的RALM进行了评估，并将其与标准RALM系统进行了比较，重点关注三个主要方面：（1）使用DPR检索的文档的整体问答性能，（2）噪声鲁棒性，通过向系统引入噪声信息进行评估，（3）未知鲁棒性，通过不在LLaMa-2预训练数据中覆盖的查询进行评估，即实时问题。实验表明，CHAIN-OF-NOTE（CON）不仅在与DPR检索的文档一起使用时提高了整体问答性能，而且在噪声和未知方面的鲁棒性也显著提高。这包括在噪声检索文档中准确性（以精确匹配分数衡量）提高了+7.9，并且在超出预训练知识范围的实时问题的拒绝率提高了+10.5。

相关工作

增强语言模型：RALMs通过结合大型语言模型和外部知识源，能够提供更具相关性和深度的答案。最近的研究工作主要集中在改进检索器和阅读器、进行端到端训练以及将检索系统与大型语言模型集成。此外，还出现了以kNN-LM为代表的新型RALMs，以及一些基于RALMs的产品，如ChatGPT插件和新版必应。鲁棒性方面的研究表明，上下文的相关性对模型性能有重要影响，而训练模型忽略不相关上下文被认为是提高RALMs效果的一个关键方向。

Chain-of-Thought（CoT）提示方法：它通过将复杂问题分解为一系列中间步骤来提高问题解决能力。CoT方法已在多模态推理、多语言场景和知识驱动应用等领域得到应用。除此之外，还有一系列其他链式方法（如解释链、知识链、验证链和信息检索链）被提出来解决LLM应用中的各种挑战。尽管这些方法在提高LLMs性能方面显示出潜力，但它们在提高检索增强型语言模型（RALMs）的鲁棒性方面的应用尚未得到充分探索。这表明在将这些策略应用于增强RALMs方面仍有进一步的研究空间。

提出方法

CON框架为检索到的文档生成顺序阅读笔记，这使得模型能够系统地评估从外部文档检索到的信息的相关性和准确性。通过创建顺序阅读笔记，模型不仅评估每个文档与查询的相关性，还识别这些文档中最关键和最可靠的信息。这个过程有助于过滤掉不相关或不够可信的内容，从而导致更准确和更符合上下文的回应。

背景知识： RALMs

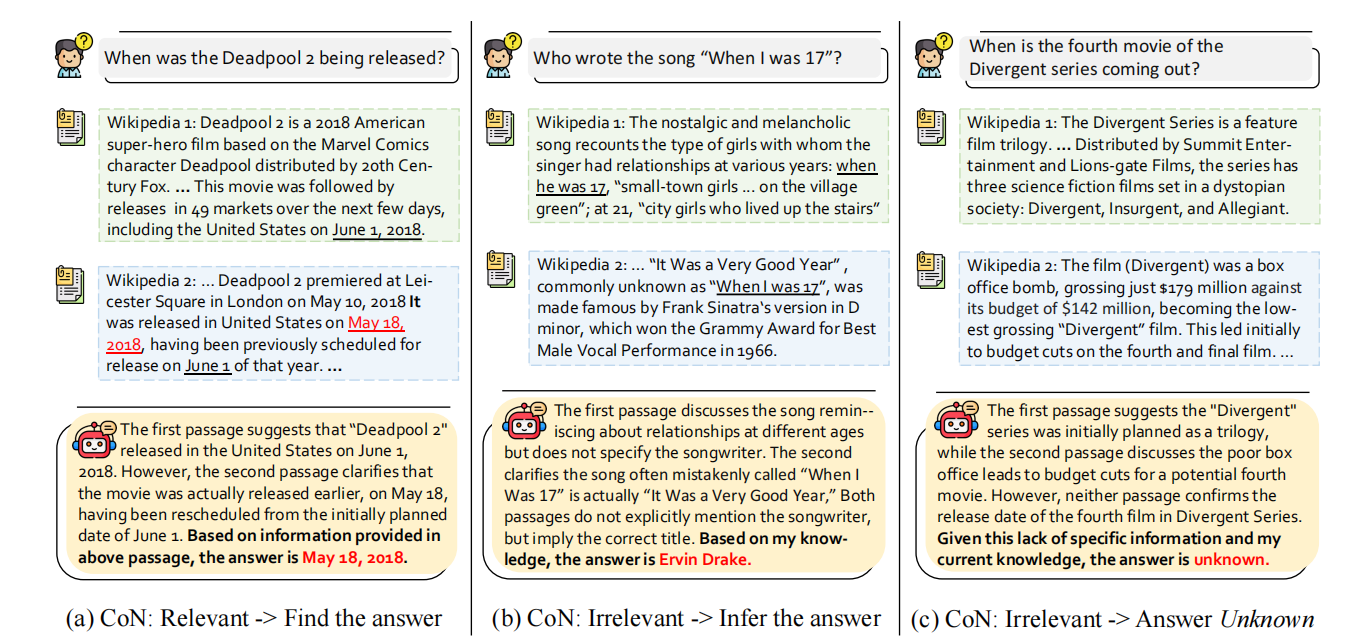
RALMs如何通过整合检索到的文档来提升语言模型的输出。这些模型通过辅助变量操作多个文档，以提供信息更丰富、更准确的响应。尽管RALMs在处理复杂问题时展现出潜力，但它们存在一些问题，例如表层处理、处理矛盾信息的困难、透明度和解释性不足，以及对检索文档的过度依赖。

CHAIN-OF-NOTE框架

CHAIN-OF-NOTE（CON）框架提出了一个解决检索增强型语言模型（RALMs）所面临挑战的解决方案。这一框架通过一个结构化的记笔记过程显著增强了RALMs评估检索到的文档的能力。具体而言，它涉及生成文档的顺序阅读笔记，使模型能够系统评估从外部文档中检索到的信息的相关性和准确性。通过创建顺序阅读笔记，CON不仅评估每个文档与查询的相关性，而且还确定了这些文档中最关键和最可靠的信息。这个过程有助于过滤掉不相关或不那么可信的内容，导致更准确和更具上下文相关的响应。CHAIN-OF-NOTE（CON）的实现包括三个关键步骤：（1）设计笔记，（2）收集数据，（3）训练模型。

3.3.1 笔记设计

这个框架主要构建了三种类型的阅读笔记，如图2所示，基于检索到的文档对输入问题的相关性：首先，当一个文档直接回答问题时，模型从这份文档中获得最终答案；其次，如果检索到的文档并不直接回答问题但提供了有用的背景，则模型利用这些信息结合其固有知识来推导出答案；最后，如果没有足够的信息来回答，模型默认回应“未知”。这种方法平衡了直接检索、推理以及对知识空白的认知之间的关系。



3.3.2 数据收集

为了使模型能够生成这样的阅读笔记，收集适当的训练数据是必要的。每个阅读笔记的手动注释是资源密集型的，因此我们采用了一种先进的语言模型ChatGPT来生成笔记数据。这种方法既经济又有效，并增强了模型的预测能力。这个过程是通过在Natural Questions（NQ）训练数据集上，指定具体指令来提示ChatGPT生成的。

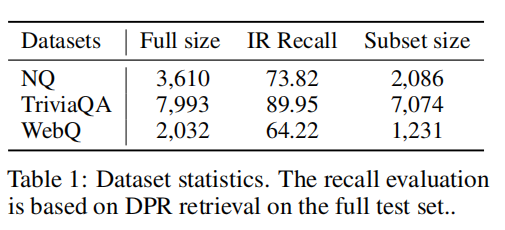
3.3.3 模型训练

收集了10K训练数据后，下一步是使用它们来训练我们的CHAIN-OF-NOTE模型，该模型基于开源的LLaMa-2 7B模型。为此，我们将指令、问题和文档作为提示连接起来，并以标准监督方式训练模型来生成笔记和答案。我们的LLaMa-2 7B内部模型学会了为每个文档顺序生成阅读笔记以评估它们对输入问题的相关性。响应是根据文档的相关性生成的，提高了准确性并减少了错误信息。对于不相关的文档，模型要么依靠固有知识给出答案，要么在答案无法确定时回应“未知”。

实验

4.1 实验设置和评估

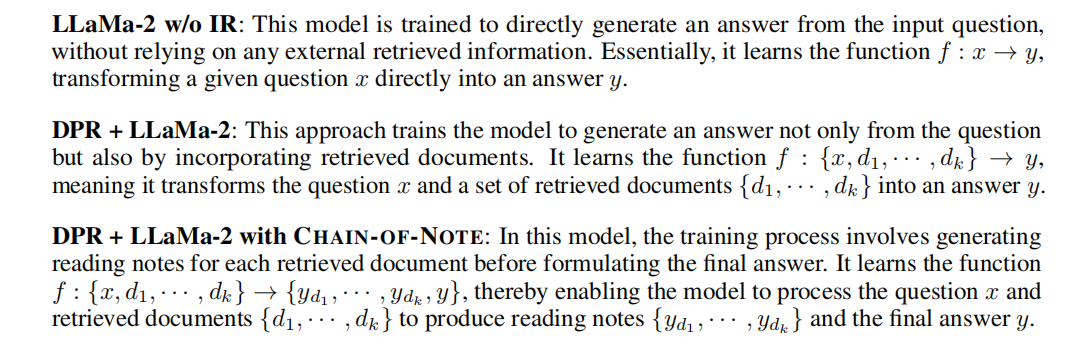
使用三个开放域问答基准数据集进行了全面的实验：NQ、TriviaQA和WebQ。此外，还使用RealTimeQA作为特殊案例来评估对“未知”情况的鲁棒性。评估基于两种评估集：完整集和子集评估。首先，类似于传统的开放域QA评估，我们使用测试集中的所有问题来评估整体QA性能。文档是使用DPR检索的，然后将top-k文档输入到生成器中。对TriviaQA的评估是在包含7993个例子的Wikipedia开发集上进行的。因此，我们也在此开发集上进行同样的评估，以便与他们的性能进行比较。其次，为了评估模型对噪声的鲁棒性和未知情况的鲁棒性，我们从上述测试集中提取了包含检索到的相关文档的子集。然后，我们枚举每一个检索到的文档，以确定它是否对于给定的问题是一个标准文档。例如，当噪声比率为20%且需要top-5文档时，则4个是相关文档，1个是无关文档。在枚举检索到的文档时，我们填充了两个列表；当其中一个列表达到其限制时，我们停止向该列表中添加更多文档，直到两个列表都完成为止。在没有检索到DPR的相关文档的情况下，我们从鲁棒性评估中排除了这些问题。因此，子集比原始测试集小，如表1所示。



4.1.1 数据集和分割

实验涉及三个开放域问答基准数据集：NQ、TriviaQA和WebQ。特别地，RealTimeQA被用来评估对未知情况的鲁棒性。两种评估集合被用来进行：一个使用完整数据集，另一个使用数据子集。

4.1.2 基线方法

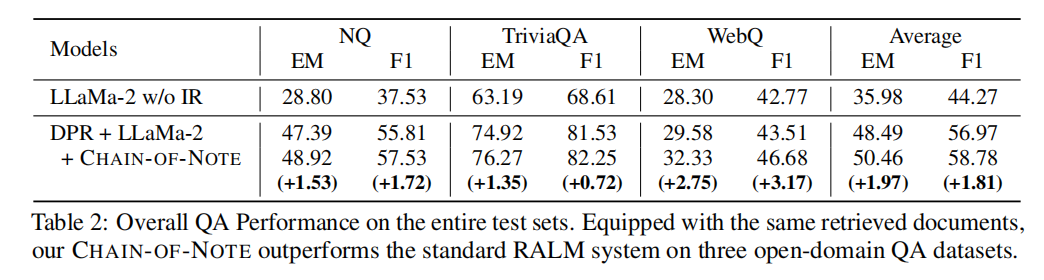


4.1.3 评估指标

对于开放域QA性能的评估，我们使用了两个广泛认可的指标：精确匹配（EM）得分和F1得分。对于EM得分，如果答案的标准化形式——通过Karpukhin等人定义的标准化程序获得——与提供列表中的任何可接受答案相符，则认为答案是正确的。与EM得分类似，F1得分将预测和正确答案视为标记包，并计算预测和正确答案之间的平均重叠。此外，我们使用拒绝率（RR）来评估当提出超出语言模型知识范围的问题时的未知鲁棒性。

4.2 整体QA性能评估

在评估中，我们将我们的方法与三个开放域QA基准测试上的各种基线进行了比较。我们注意到，具有检索功能的RALM（DPR + LLaMa-2）一致性地优于没有检索的LLaMa-2。这一改进与检索过程的有效性密切相关。如表中所示，DPR在NQ和TriviaQA数据集上展示了明显优越的检索性能，相比之下WebQ的性能较弱。因此，检索的好处在NQ和TriviaQA上更为明显。



4.3 噪声鲁棒性评估

我们的噪声鲁棒性评估在两种场景下进行：使用通过检索获得的噪声文档（通过去除检索集中的相关文档并保留排名最高的不相关文档）和使用从整个Wikipedia随机抽取的完全随机文档。噪声检索文档通常包含误导信息，因为它们在语义上与输入问题相似，这与代表完全噪声的随机文档形成对比，显示了CON增强的RALM比标准模型有更好的性能。

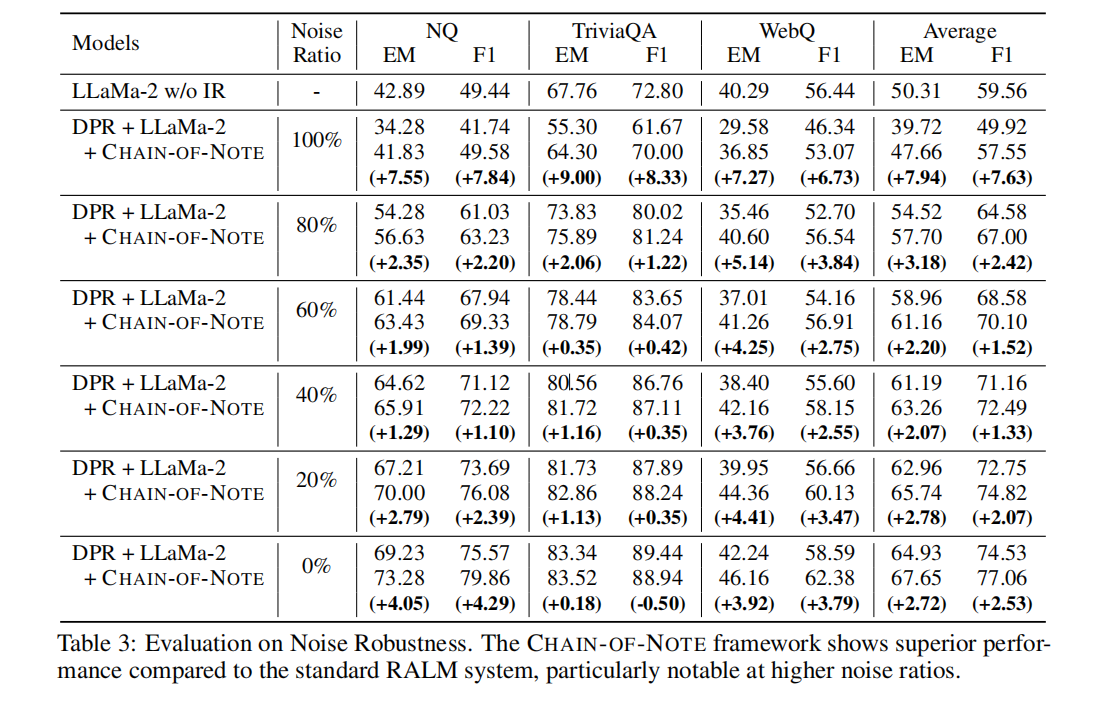
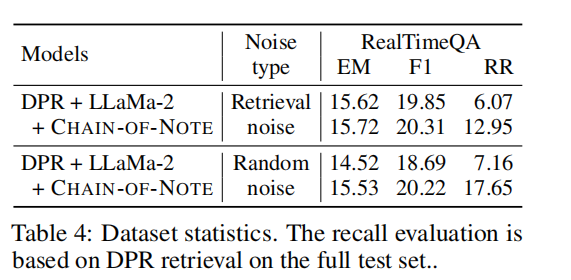
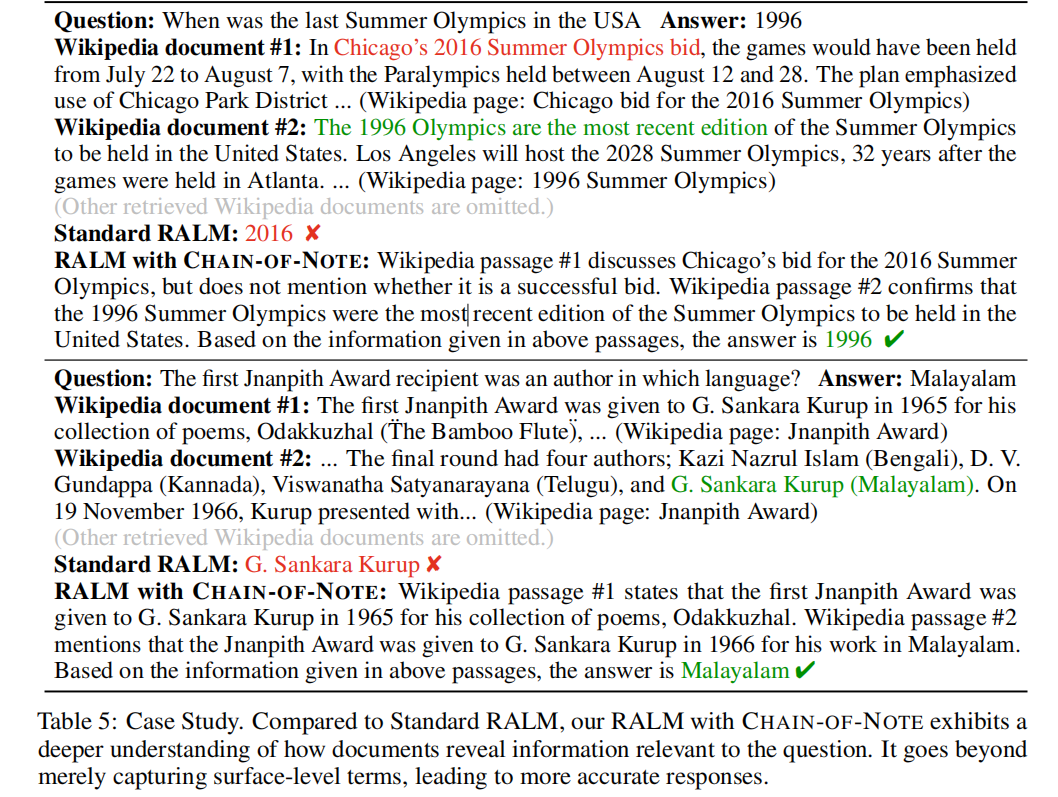
4.4 未知鲁棒性评估

表4说明了我们配备了CON的RALM在处理未知情况方面显示出更高的鲁棒性，特别是在RealTimeQA基准测试中，该测试不在模型的域内，也不是预训练数据的一部分。展示了装备CON的RALM在处理初始训练期间未涵盖的信息的查询方面有更好的能力。



4.5 案例研究

在我们的案例研究中，如表5所示，我们比较了标准RALM和我们增强的RALM与CON的响应。这些示例突出了每个模型处理和解读检索文档的信息的差异，说明了CON框架在分析和处理检索信息时比标准RALM方法更为精确和有效。



总结

在本文中，我们介绍了CHAIN-OF-NOTING（CON）框架，这是一种旨在增强RALMs鲁棒性的新方法。CON的核心概念是为每个检索到的文档生成顺序阅读笔记。这个过程允许深入评估文档与所提出问题的相关性，并帮助综合这些信息来构建最终答案。我们利用ChatGPT生成了CON的初始训练数据，随后使用LLaMa-2 7B模型进行了进一步的完善。我们在各种开放域QA基准测试中进行的测试显示，与CON集成的RALMs在性能上大幅超越了传统RALMs。