EnronQA：在私人文档上实现个性化RAG

[MichaelJ. 瑞安](https://orcid.org/0000-0002-4353-821X)

美国加利福尼亚州帕洛阿尔托

徐丹梅

雪花

美国加利福尼亚州圣马特奥

克里斯·尼维拉

雪花

美国加利福尼亚州圣马特奥

丹尼尔·坎波斯

雪花

美国加利福尼亚州圣马特奥

多用户检索

填写2000年年底业绩管理过程的反馈表的截止日期是什么

搜索8429电子邮件

103，638封电子邮件; 528，304个QA对; 150个用户

二○年十一月十七日（星期五）

图1：EnronQA基准测试在超过100，000封电子邮件的清理语料库上实现了个性化和私人检索基准测试，这些电子邮件跨越了150个用户的528，304个优质问答对。我们探索单用户和多用户检索设置。

# 摘要

检索增强生成（RAG）已成为将知识密集型上下文引入大型语言模型（LLM）的最流行方法之一，因为它能够在推理时引入本地上下文，而不会产生与微调相关的成本或数据泄漏风险­。 私人信息与LLM培训的明确分离使RAG成为许多企业LLM工作负载的基础，因为它允许公司­使用客户的私人文档来增强LLM的理解。尽管RAG在企业部署中的私有文档很受欢迎，但当前用于验证和优化RAG管道的RAG基准测试得出了它们的

在Snowflake实习期间完成的工作电子邮件地址：[michaeljryan@stanford.edu](mailto:michaeljryan@stanford.edu)

[本作品是在知识共享署名4.0国际许可协议下授权的。](https://creativecommons.org/licenses/by/4.0/legalcode)

语料库来自公共数据，如维基百科或通用网页，并提供很少或根本没有个人背景。 为了增强RAG的个性化和私密性，我们发布了EnronQA基准测试，这是一个包含103，638封电子邮件的数据集，其中包含150个不同用户收件箱中的528，304个问答对。EnronQA能够更好地对RAG管道进行私人数据基准测试，并允许在实际数据上引入个性化检索设置的实验。最后，我们使用EnronQA探索权衡记忆和检索推理时，私人文件。1

1. 介绍

检索越来越成为­在称为检索增强生成（RAG）的过程中向LLM添加上下文的最常见方式之一[[20](#_bookmark37)，[22](#_bookmark39)，[43](#_bookmark60)]。RAG管道涉及增加自然

1在此Huggingface repo上发布的所有数据：[MichaelR 207/enron\_qa\_0922](https://huggingface.co/datasets/MichaelR207/enron_qa_0922)

语言生成能力的LLM与外部数据存储。 这种增强通过在文档中生成来提高LLM的真实性[5，66]和可解释性[68，75]。此外，RAG已被证明可以帮助LLM通过检索相关知识来解决知识密集型任务[ 36 ]，而不是依赖LLM来记忆事实。 RAG最受欢迎的应用之一是对私人文档的检索，­使公司和用户能够与大量的内部和私人知识库进行交互[[14，](#_bookmark31)[28]。](#_bookmark45)

虽然对私有文档的检索是RAG最流行的用例之一，但相对而言，很少有大规模的RAG基准测试关注私有文档检索[[4](#_bookmark21)]。 用于验证和优化RAG管道的最流行的基准测试从维基百科[32，41，77]或公共互联网[[6](#_bookmark23)]中获取语料库。 我们将在2.1节中进一步讨论这一点[。](#_bookmark2)

此外，最近人们对隐私保护RAG[[78](#_bookmark95)，[79](#_bookmark96)]感兴趣，其中模型能够在不暴露个人身份信息的情况下帮助和访问私人文档中的知识是很重要­的。基于私有文档的个性化RAG­[21，72，80]和联合信息检索[58，58，65，71]需要个人文档和分段数据存储。这样的探索­将受益于一个现实的语料库分割成几个私人用户开发私人和个性化的RAG测量方法。

为了服务于这些不同的任务，并增加私人环境中基准标记的覆盖范围­，我们引入了基于安然电子邮件语料库的EnronQA基准[[17](#_bookmark34)]。 EnronQA包含103，638封电子邮件，其中包含528，304个问答对，跨越150个不同的用户收件箱。 通过设计一个严格的问题生成管道，以特定的评估为基础，我们确保收集高质量和多样化的问题。 我们的QA数据集在大小上对于私有文档的公开可用文档检索是无与伦比的，并且足够大，可以在此设置上进行微调，优化和评估。图[1](#_bookmark0)展示了EnronQA基准测试和我们测试的一些评估设置。

为了展示EnronQA的实用性，我们进行了两个案例研究。

首先，我们展示了如何基准RAG管道上安然有更高的净空，以提高检索质量（**？？**).我们发现，没有检索，RAG管道得分低于5%的EnronQA，不像其他流行的RAG基准，使用参数LLM知识，它有可能得分超过60%，而无需任何重新­trieval。接下来，我们训练LoRA适配器来记忆我们大型数据集中的事实知识。 我们的LoRA记忆适配器­显示，训练LLM记忆私人事实­知识可以与在上下文中存储所有事实相提并论，但是检索特定的相关信息仍然优于两者。

总体而言，我们的贡献如下：

1. EnronQA基准测试，收集了超过100，000封私人­电子邮件和500，000个问题，分为150个­不同的用户收件箱[（§3）。](#_bookmark4)
2. 我们展示了我们的基准测试的质量和实用性，并将其与其他流行的RAG基准测试进行比较[（§4.2）。](#_bookmark12) 我们在EnronQA上对流行的RAG管道进行基准测试，作为未来工作的[基准（§5）。](#_bookmark13)
3. 我们激励记忆私人知识，并展示­了一种基于LoRA的记忆事实知识的方法­，这种方法与将知识库置于上下文中的方法相比具有竞争力[（§6）。](#_bookmark16) 检索最相关的信息­优于两者，但我们讨论未来的­改进，激励进一步探索记忆适配器。

# 相关工作

我们组织相关工作的讨论，以跨越我们的核心贡献­：基准RAG，并在LLM事实记忆

# 检索增强生成基准

我们在表1中提供了一个流行的RAGQA基准的简短列表，在表[6中提供了一个更全面的列表。](#_bookmark100) 几个常见的RAG基准从维基百科中提取文档[[2](#_bookmark19)，[33](#_bookmark50)，[35](#_bookmark52)，[41](#_bookmark58)，[74](#_bookmark91)，[77](#_bookmark94)]。不幸的是，这使得基准测试不太适合使用现代LLM对RAG管道进行基准­测试，这些LLM已经记住了维基百科和一般知识的大量内容[[57]。](#_bookmark74)

与我们最相关的工作是ConcurrentQA[[4]，](#_bookmark21)它创建了一个依赖于安然电子邮件语料库和维基百科的多跳推理的基准我们认为这是一个很好的资源与我们的结合，但值得注意的是，这些基准解决不同的问题。首先，ConcurrentQA基准测试仅限于一个收件箱，而EnronQA跨越150个不同的用户，从而可以在更大的基准测试中研究个性化的RAG。其次，ConcurrentQA有18.4k个QA对，而不是EnronQA中的528.3k个，这使得我们的­基准更适合于涉及微调事实知识和持续预训练的探索。 EnronQA将使­这些新兴的信息检索趋势的探索。最后，ConcurrentQA由多跳查询组成，这对于测试复杂的管道非常有用。相反，我们将EnronQA设计为单跳，以专注于每­声化和记忆的有趣案例，但是在第3.1节中，我们讨论了如何使EnronQA与ConcurrentQA完全兼容，以便用户可以使用我们的基准测试单跳和多跳检索。

# 事实化

在我们的案例研究中，我们探讨了事实记忆作为替代传统的检索回忆事实。LLM中的事实记忆是一个令人兴奋且相对较新的方向。LLM记忆的大部分工作来自LLM中学习遗忘的­工作[45，48]或从­互操作性透镜中理解记忆[[26](#_bookmark43)]。最近的一些作品已经期待在LLM鼓励记忆的策略。一种方法是用外部存储器参数扩充LLM [[12，23](#_bookmark40)]。其他工作鼓励对LLM进行微调，以实现事实记忆[46，62，69]。­

也许最相关和最令人兴奋的连接，特别是信息检索是持续的预培训[[24](#_bookmark41)，[37](#_bookmark54)]，其中LLM适应新的知识和领域，有可能只是继续对来自新来源的更多文档进行预培训。最有希望的连接到我们的设置是“合成

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 基准 | 文件 | QA对 | 域 | 源 |
| ConcurrentQA[[4]](#_bookmark21) | 5.2百万+47千 | 18.4k | 一般知识和私人知识 | 维基百科+电子邮件 |
| 电子手册[[54]](#_bookmark71) | 308k | 3.3k | 客户支持 | TVManual |
| 哈格里克[[35]](#_bookmark52) | 32.8M | 2.6k | 一般知识 | 维基百科 |
| [[第77话]](#_bookmark94) | 5.2M | 112.8k | 一般知识 | 维基百科 |
| INSCIT[[74]](#_bookmark91) | 6.6M | 4.7k | 一般知识 | 维基百科 |
| 马可女士[[6]](#_bookmark23) | 3.6M | 1.01M | 一般知识 | 网页 |
| 自然问题[[41]](#_bookmark58) | 5.9M | 323K | 一般知识 | 维基百科 |
| [[31]](#_bookmark48)第三十一话 | 211.3k | 273.5k | 学术研究 | 研究摘要 |
| QReCC[[3]](#_bookmark20) | 10M | 81k | 一般知识 | 网页 |
| 搜索QA[[16]](#_bookmark33) | 6.9M | 140.5k | 一般知识 | 谷歌搜索 |
| 技术质量保证[[8]](#_bookmark25) | 802k | 1.4k | 客户支持 | 技术论坛 |
| TopiOCQA[[2]](#_bookmark19) | 5.9M | 50k | 一般知识 | 维基百科 |
| [[33]](#_bookmark50)第三十三话 | 662.7k | 96k | 一般知识 | 维基百科 |
| EnronQA（我们的） | 103.6k | 528.3k | 私人知识 | 电子邮件 |

## 表1：基于文档的QA基准比较。 我们仅限制语料库大小超过50 k文档的资源空间，但在附录A中提供了完整的比较[。](#_bookmark99) EnronQA涵盖了与许多流行的QA基准相当的语料库规模，同时拥有更多的QA对，可以进行训练，优化和文档记忆探索。此外，EnronQA使用电子邮件跨越了未开发的私人文档领域

Continued Pretraining”[[76](#_bookmark93)]，其中实体从文档中提取，并在这些实体之间绘制连接然后，LLM在这些连接上持续进行预训练。在这样做的过程中，作者发现，这鼓励了对文档的记忆，最终，当用RAG进行测试时，性能效益是复合的。在这项工作中，我们发布了一个基准的私人和unmemorized文件基准RAG性能。这样的资源将是一个丰富的测试平台，为持续的预训练方法提供现实的QA任务

# EnronQA数据集构建

我们使用Enron电子邮件­核心构建EnronQA基准[[39](#_bookmark56)]。 联邦能源监管委员会在2003年西部能源市场调查期间发布了原始语料库。 原始数据集包含超过60万条消息和158个不同的用户（收件箱）。 我们使用2015年发布的语料库，该语料库已被清理，并应安然参与者的要求删除了电子邮件[[17](#_bookmark34)]。这个版本的语料库包含150个用户的517，401封电子邮件。 为了将原始电子邮件转换为高质量的RAG基准，我们设计了一个3阶段的管道：过滤[（§3.1），](#_bookmark5)QA生成[（§3.2），](#_bookmark7)后处理[（§3.3）。](#_bookmark10)

# 语料过滤

对于我们的数据过滤器，我们从流行的预训练数据过滤管道中汲取灵感[13，19，59，67]。 从原始电子邮件中过滤数据与清理非结构化Web数据有许多相似之处。表[2](#_bookmark6)概述了过滤过程的每个步骤，以及删除的电子邮件数量和删除的电子邮件主题示例

*- 是的* Web数据提取管道使用minhash去­重[[42](#_bookmark59)]来减少相同文档的数量，并以最小的编辑（如软件许可证）来减少文档的数量。在电子邮件收件箱中，这可以与电子邮件线程和电子邮件订阅相对应。例如，订阅天气服务可能每天发送相同的电子邮件，但天气预报会发生变化。我们修改了text-dedup [[52](#_bookmark69)]库的minhash实现，增加了最后一步，计算匹配文档之间的Jaccard相似度，我们使用9个波段和27行，以0.9的Jaccard相似度阈值运行minhash去重为了处理线程重复数据删除，我们还删除了在另一封电子邮件的内容中完整出现的任何电子邮件，我们称之为“子集”重复数据删除。

*Gopher质量过滤器。* [[59](#_bookmark76)]概述了一些计算来过滤预训练文档的质量。它们根据文档单词长度、平均单词长度、以省略号结尾的行数以及字母数字字符与符号的比率进行过滤。为我们的电子邮件域调整这些规则的截止值，我们将截止值设置为50到1000个单词之间的电子邮件，长度在3到10之间，字母数字字符与符号的比例大于或等于0.65，并且以省略号结尾的行少于10%。这有助于过滤掉太短的电子邮件和过长或低质量的电子邮件，例如软件系统的自动日志文件和自动财务报告。

*语言识别。安然员工之间的*一些电子邮件线程­发生在非英语语言。由于我们使用有限多语言培训的LLM创建英语基准，因此我们过滤了非英语或分类文档

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 步骤 | 操作 | 文件 | 变化 | 受影响电子邮件的主题示例 |
| 0 | （原件） | 517,401 | – | – |
| 1 | MinhashDedup | 228,098 | -289 303人 | “每日图表和矩阵作为热链接5/15”，“安然提及-05/12/01-05/13/01” |
| 2 | 子集重复数据消除 | 192,759 | -35 339人 | “FW：FW：ResearchLibrary”，“Re：FW：Fedex”，“Pre-Party？？“、“Re：Reference” |
| 3 | 质量过滤器 | 103,513 | -89 246人 | “回复：606925 Finney Attasseril SRF（mar）"，“EnronOnline版权声明” |
| 4 | 语言ID | 103,164 | -三百四十九 | “Praca dyplomowa”，“Saludos，Feicitaciones y Follow up." |
| 5 | NSFW滤波器 | 103,015 | -一百四十九 | 你想和我玩吗？“，“回复：理查德糟透了！" |
| 6 | 毒性过滤器 | 102,603 | -四一二 | “Re：LSU/Florida”，“FW：Isn't it Funny”，“I AM FEELING NEGLECTED” |
| 7 | 并行QA | 103,638 | +1，035 | “2001年11月16日互联网日报”，“加州能源危机时间轴” |

## 表2：贯穿语料库过滤的几个步骤的安然电子邮件语料库以及在每个步骤添加/移除的来自电子邮件的主题的样本

使用fastText语言识别模型，英语的置信度低于80%[[34]。](#_bookmark51)

*NSFW/毒性过滤器。令*人惊讶的是，我们发现一些有毒和不适当的内容作为一些电子邮件的主题。为了保护参与者的隐私并保持我们EnronQA资源的专业性，我们会过滤掉这些文档。我们使用两个在jigsaw [ 1 ]上训练的fastText分类器用于dolma项目[[67](#_bookmark84)]，一个用于检测毒性，一个用于检测NSFW含量。我们会过滤掉那些不被预测为安全的电子邮件，置信度大于或等于

*并发质量保证。为了*使我们的资源与相关的ConcurrentQA[[4](#_bookmark21)]基准兼容，我们将ConcurrentQAEnron语料库中的所有文档映射回我们语料库中的电子邮件。如果这些规则过滤掉了任何一个，我们就把它们加回去。

# QA生成管道

我们必须生成高质量的问题，以将清理后的语料库转换为RAG基准。我们设计了一个多级复合LLM系统，在DSPy中实现并优化了端到端[[38](#_bookmark55)]。一个问题的生成包括10到50个不同的LLM调用，每个调用都被设计用于单个模块的目的。我们的流水线在图[2](#_bookmark8)中进行了直观的描述，可以概括为4个主要阶段：（1）初始生成，（2）评估，（3）反馈生成，（4）优化。­

*第一代。*我们生成一个初始问题，给定一个文档和该文档的一组先前问题（以便LLM不会生成重复）。 我们的问题生成器模型是llama3.1- 70 b [[15](#_bookmark32)]，提示符由DSPy使用MIPROv 2优化器优化[[55](#_bookmark72)]。 我们的优化目标是减少早期停止管道的细化步骤的数量。 如果下面描述的任何*评估*步骤失败，则管道将重复。我们优化的提示将管道从平均1.94次重复降低到1.64次重复。我们在附录B.1中包含了我们的初始和优化提示[。](#_bookmark101)

*评价*给定一个问题，我们要评估这个问题是否是高质量的。为此，我们提出了四个评价标准­和具体措施：具体性、客观性、基础性和质量。

**的特异性**我们指定一个问题为“具体”，如果有十个类似的电子邮件（包括真正的电子邮件的问题是关于）一个法学硕士可以挑选出哪封电子邮件将回答这个问题。我们通过从我们的语料库中检索给定问题的前10个相关文档来挖掘困难的负面例子 我们使用构建在Snowflake arctic-embed-m-v1.5[[50](#_bookmark67)]上的biencoder来检索前10个最相似的文档­。 我们使用Llama3.170b作为我们的选择器LLM。 完整提示见附录[B.2。](#_bookmark102)

**客观。**我们确定一个问题是“客观的”，如果两个模型从不同的家庭回答相同的问题-

同样的答案，考虑到电子邮件作为上下文。这里我们使用Llama3.1 70b指令[[15](#_bookmark32)]和Mixtral 8x7B指令[[30](#_bookmark47)]。 我们使用LLM作为判断，以确定答案是否匹配。我们的Llama3.170B指令LLM法官实现了

0.98 F1评分，在200代小型研究中进行人类评价我们在附录中包括所有QA提示

[B.5](#_bookmark105)和我们的法学硕士作为法官评估的细节在­Appen [B.6。](#_bookmark106) 鉴于该问题被认为是客观的，我们将Llama 3.1 70 B Instruct输出保存为“黄金答案”。“**接地。**我们确定一个问题是“接地”，如果骆驼或混合模型可以回答相同的黄金答案没有电子邮件作为上下文。这两个测试，我们的问题的答案没有记住，问题是不容易猜测。同样，我们使用LLM作为评判，以电子邮件为基础，以确定答案是否与从之前的评估中获得的“黄金答案”相匹配。如果没有一个不接地的答案与黄金答案相匹配，我们认为这个问题是“接地”的。“所有QA提示均包含在附录B.5中[。](#_bookmark105)

**质量.** 我们最后的评估步骤是衡量“质量”

由法学硕士法官根据人类的判断来回答这个问题我们使用管道生成了20个问题，只有具体性，客观性和隐蔽性阶段。我们让两位作者根据评估特异性、客观性和隐蔽性的标准将它们标记为“高”、“中”或“低”质量 作者的注释具有0.5 Spearman相关­性，第三作者裁定了分歧。 同一作者还独立地将另外21个问题标记为“高”，“中”或“低”。使用21个单标签问题作为开发集

## 图2：我们的多级复合LLM系统在安然电子邮件语料库上的QA生成。我们的流水线由上图中标记的4个阶段组成：（1）初始生成，（2）评估，（3）反馈生成和（4）优化。生成一个高质量的问题需要10-50个不同的LLM调用，并且系统是端到端优化的 我们的管道声称，问题是具体的，客观的，有根据的，高质量的（与人类的判断相关）。所有Llama图标都对应于Llama 3.1 70 b指令[[15](#_bookmark32)]，Mistral图标代表Mixtral-7 B-Instruct模型[[30](#_bookmark47)]，我们的检索器是使用Snowflake的artic-embed-m-v1.5的双向编码器[[50]。](#_bookmark67)

和20组标记的问题作为测试集，我们设计了一系列规则，用于Llama-3.1 70 B指令来确定问题是“高”还是“低”质量（将“中等”标签包装成低质量）。 我们的规则使法官在开发集上达到85.7%的准确率，并在测试集上运行一次，产生了85%的准确率。我们的评估管道的最后阶段使用了一个Llama3.170B指令模型，并使用了一个规则来确定问题是否是“高质量”的。“我们在附录[B.7中包括了我们的规则。](#_bookmark107)

*反馈生成。*根据问题在评估阶段所处的最新阶段，我们生成反馈以添加到细化步骤的上下文中。

* 如果问题不是特定的，我们将在“优化”步骤中描述的特殊情况下处理它
* 如果问题不客观，我们会提供反馈：“问题不客观。 不同的注释者在给定相同的电子邮件作为上下文的情况下以不同的方式回答相同的问题。可能会受益于更清晰。"
* 如果问题不接地，我们提供反馈：“问题不接地。没有阅读电子邮件就很容易猜到这个问题的答案。"
* 如果问题没有通过质量检查，我们使用LLM的思维链推理作为为什么问题质量不高的反馈。这通常会引用电子邮件失败的

*精炼。*如果我们的问题在所有评估阶段都成功，它将被认为是一个好问题，并添加到我们的题库中。否则，我们需要改进它。 我们有两个改进步骤。如果

## 图3：问题重写管道。首先，我们要求Llama 3.1 70B指令重写问题，然后我们要求它回答这个新问题。最后，我们使用Llama 3.1 70B In结构来检查答案是否匹配。

如果问题不具体，我们向LLM显示从具体性检查中检索到的十封电子邮件，并要求它重写问题，使其更具体地仅针对黄金电子邮件。如果问题在其他步骤中失败，我们使用生成的反馈要求LLM重写问题并解决反馈。特异性和一般反馈重写提示都使用DSPy进行了优化。我们将初始和优化的特异性细化提示纳入附录

以及附录B.4中的初始和优化反馈提示[。](#_bookmark104)

# 附加数据处理

我们提供了问题的重写版本，以使我们的数据集对各种下游任务更实用，例如我们的备忘­录化案例研究[（第6节）。](#_bookmark16)在训练LLM记住特定信息的情况下，这使您能够在保留信息内容的同时训练和测试不同的问题。 在图[3中，](#_bookmark9)我们展示了重新表述问题的管道。我们使用Llama3.1-70 B-Instruct重写问题，回答重写的问题，

最后判断答案是否相同。 如果答案不匹配，我们会再试最多15次，然后放弃这个问题。 我们在这个过程中丢弃了265/528，569个问题。

除了我们数据集的核心组件：问题，黄金答案，电子邮件和重新措辞的问题，我们还发布了­创建核心数据集时产生的错误工件。这些工件包括来自评估步骤的Mixtral-8x 7 B-Instruct模型的验证答案，以及­Llama3.1- 70 b-Instruct和Mixtral-8x 7 B-Instruct在回答EnronQA基准测试中以Oracle文档为条件的每个问题时的思维链。

# 数据集质量

我们在这里讨论EnronQA基准的一些属性，以及是什么使它成为社区的高质量和有价值的资源。

# 数据集统计数据

我们在表3中报告了EnronQA基准的汇总统计数据[。](#_bookmark11)值得注意的是，基准测试包含超过333k个训练问题，关于单个用户电子邮件的问题的中位数超过1k。EnronQA基准测试对于问题的微调、文档的持续预训练和­RAG管道的基准测试来说是相当大的。

|  |  |  |  |
| --- | --- | --- | --- |
| 度量 | 火车 | Dev | 测试 |
| 电子邮件统计数据（跨拆分一致） | | | |
| 每个用户的平均电子邮件数 |  | 491.81 |  |
| 每位用户的邮件数中位数 |  | 240.5 |  |
| 平均电子邮件长度（字符） |  | 2,269.69 |  |
| 电子邮件长度中位数（字符） |  | 1,664.0 |  |
| 问题统计 |  |  |  |
| 每个用户的平均问题数 | 2,223.15 | 703.43 | 595.44 |
| 每位用户的问题中位数 | 1,063.0 | 334.5 | 278.0 |
| 每封电子邮件的平均问题数 | 4.52 | 1.43 | 1.21 |
| 每封电子邮件的中位数问题 | 3.0 | 1.0 | 1.0 |
| 问题总数 | 333,473 | 105,515 | 89,316 |

表3：EnronQA基准的汇总统计。基准­包含适当数量的文档和问题，用于继续相关和RAG基准。

* 1. 校准

使用公共RAG基准的一个缺点是，从维基百科提取文档，缺乏基准­分数和检索质量之间的校准­EnronQA的一个比较优势是，在大多数情况下，LLM中编码的参数知识­不会记住安然电子邮件。因此，你可以期待从一个更好的寻回犬的收益，以匹配基准的准确性收益为了验证这一假设，我们选择了两个标准的RAG基准**NaturalQuestions**[[41](#_bookmark58)]和**TriviaQA**[[32](#_bookmark49)]。 NaturalQuestions包含323，000个Google查询，答案涵盖590万个维基百科文档。TriviaQA包含95K问答对，由琐事爱好者撰写

60万篇文章我们特别使用KILT[[56](#_bookmark73)]版本的数据集。

*实验设置。我们*为NaturalQuestions和TriviaQA过滤了10，000个训练/ 500个验证示例。为EnronQA提供1，000个培训/ 500个验证示例 我们使用NaturalQuestions和TriviaQA的完整验证集作为测试集。 对于EnronQA，我们使用实际测试集。 我们使用MIPROv2优化[[55](#_bookmark72)]为每个设置优化两个DSPy程序。第一个程序没有上下文，必须直接回答问题第二个程序以黄金文档为背景，回答问题­。 我们优化与Llama-3.1-8B-Instruct作为我们的任务模型和Llama-3.1- 70 B-Instruct作为我们的提示模型与10个候选程序。 在运行实验时，我们使用Llama-3.1- 70 B-Instruct。我们运行无上下文的情况;然后，我们通过随机包含正确的文档作为上下文或随机抽样的文档来模拟Recall@1在0.0和1.0之间。 我们用五个随机种子进行测试，并对结果进行平均。­使用Llama-3.1- 70 B-Instruct作为与黄金答案相比的答案准确性的评判

*结果*我们将实验结果包括在图[4中。](#_bookmark14) 我们发现，EnronQA是唯一的基准添加上下文总是优于无上下文的基线。 对于NaturalQuestions，它需要一个Recall@1高于0.5的检索器，以便超越无上下文基线。 同样，在TriviaQA上，问题更严重。 直接问Llama-3.1- 70 B-Instruct问题，不需要上下文，比所有基于检索的系统都要好，Recall@1小于0.85！这意味着在TriviaQA上对RAG管道进行基准测试的任何准确性变化（Recall@1小于0.85）可能更多地与LLM的记忆知识有关，而不是检索质量。

相比之下，EnronQA的所有改进都更准确，

上下文直接导致更高的准确性的基准。 此外­，EnronQA展示了每一点召回增益的准确性的最高改进。回忆率每提高1%，准确率就提高近0.6%­。这是因为EnronQA中的知识没有被大型基础模型记住，这是使NaturalQuestions和TriviaQA变得平凡化的问题

# RAG管道基准

为了提供一些基准性能数据并展示EnronQA在RAG基准测试中的实用性，我们测试了两种流行的检索器，三种流行的LLM和两种常见的RAG管道架构。

# 实验环境

*猎犬* 我们使用PySerini实现[[44](#_bookmark61)]和**ColBERTv2** [[64](#_bookmark81)]在完整的103，638封电子邮件中测试BM 25。对于检索器的一次调用，我们同时检索五个文档。对于每只寻回犬，我们还报告召回@5。

*大型语言模型。* 我们使用**Llama-3.1-8B-Instruct**、**Llama-3.1- 70 B-Instruct**[[15](#_bookmark32)]和**GPT 4 o进行测试，**以测试不同­尺度和家族的模型。

*RAGArchitectures.* 我们测试两个RAG设置;首先，我们测试**NoQuery Rewrite**，在这里我们直接使用

自然问题

TriviaQA

EnronQA

80

70

60

50

40

30

20

10

0.0 0.2 0.4 0.6 0.8 1.0

模拟召回@1

0.0 0.2 0.4 0.6 0.8 1.0

模拟召回@1

0.0 0.2 0.4 0.6 0.8 1.0

模拟召回@1

图4：校准实验结果。尽管所有的基准测试都大致线性地扩展更准确的上下文，但EnronQA是唯一一个添加上下文总是优于无上下文基线的基准测试对于TriviaQA，Recall@1需要接近0.85才能超过无上下文基线的性能。

无查询重写（acc）查询重写（acc）查询重写（R@5）

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | R@5 | 美洲驼8B | 美洲驼70b | GPT4o | 美洲驼8B | 美洲驼70b | GPT4o | 美洲驼8B | 美洲驼70b | GPT4o |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| BM25 | 87.5 | 72.1 | 77.1 | 81.2 | 66.1 | 80.0 | 81.1 | 82.5 | 86.6 | 88.5 |
| 科尔伯特 | 54.1 | 65.8 | 74.6 | 74.9 | 56.3 | 69.1 | 73.3 | 54.1 | 59.4 | 59.3 |

## 表4：在EnronQA基准测试中对几种检索方法和LLM进行基准测试，包括查询重写和不重写。令人惊讶的是，简单的检索基线（BM 25）在我们的基准上工作得很好。这可能是由于电子邮件中的查询和专有名词（如名称和事件）之间存在一些词汇重叠。

直接从EnronQA提问。然后，我们提供检索器的前5个结果，并将检索到的文档和问题传递给LLM进行回答。我们还测试了**查询重写**设置，我们首先让LLM将问题重写为搜索查询。然后，我们检索了五封电子邮件。最后，给出五封电子邮件和问题，我们让LLM给出答案。 之前的工作已经发现使用LLM重写查询以帮助适应特定检索器的细节[[47](#_bookmark64)]，因此我们在我们的基准测试中对此进行了测试。对于这两种设置以及所有模型和回收器，我们使用DSPy MIPROv2[[55]](#_bookmark72)优化提示和少量演示，其中有10个候选项和20个试验。

* 1. 结果

我们在表[4中给出了结果。](#_bookmark15) 我们从BM 25检索器中发现了令人惊讶的高准确性，拥有87.5的Recall@5，无需任何额外的查询重写步骤。这可能是由于­某些查询和电子邮件内容之间的词汇重叠。 因为我们的问题生成管道被优化为足够具体，可以从一批10封电子邮件中选择一封，所以查询必须命名电子邮件中的特定实体。 这反映在BM 25的高精度上。我们发现，毫不奇怪，更大的模型在我们的基准测试中表现更好，性能从8b扩展到70b再到GPT4o。我们还发现查询重写并不是特别重要

对这个基准测试很有帮助，特别是对于BM 25。最高的每形成设置是GPTO，使用BM 25进行查询重写和不进行查询重写，在EnronQA上达到了81.2%的准确率。

# 案例研究：知识化

随着越来越多的关于持续预训练的文献[[76](#_bookmark93)]，我们注意到我们的基准测试的一个有趣的用例是用于持续预训练记忆的大规模和现实的测试床。由于我们的基准包含LLM没有经过大量预训练的私人知识，以及超过50万个问题和答案对，因此有大量的数据可以进行基准测试，甚至微调模型来测试参数知识记忆，并将其与RAG进行基准测试。

为此，我们在这个方向上提供了初步的结果，希望这个资源将有助于未来的研究人员探索持续的预训练和记忆与LLM。

# 实验环境

我们想通过三种机制来探索关于文档的10到20，000个事实的记忆/检索：**长上下文**，**RAG**和**重复**。对于这种设置，我们通过直接查看问答对而不是文档来简化问题，尽管我们希望未来的工作也可以探索对文档的培训。我们使用重新措辞的问题和答案对

#事实↓ 长上下文 RAG 记忆（LoRA）

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 排名→ | - | - | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 | 2048 |
| 10 | 0.80 | 1.00 | 0.80 | 0.80 | 0.80 | 0.80 | 0.80 | 0.80 | 0.90 | 0.80 | 0.80 |
| 100 | 0.91 | 0.95 | 0.76 | 0.76 | 0.80 | 0.84 | 0.83 | 0.85 | 0.87 | 0.88 | 0.78 |
| 500 | 0.83 | 0.91 | 0.75 | 0.73 | 0.79 | 0.81 | 0.78 | 0.82 | 0.79 | 0.80 | 0.80 |
| 1000 | 0.79 | 0.89 | 0.72 | 0.71 | 0.79 | 0.79 | 0.79 | 0.78 | 0.76 | 0.78 | 0.73 |
| 5000 | NA | 0.92 | 0.53 | 0.55 | 0.69 | 0.75 | 0.74 | 0.69 | 0.78 | 0.03 | - |
| 10000 | NA | 0.92 | 0.53 | 0.61 | 0.69 | 0.74 | 0.77 | 0.69 | 0.78 | 0.79 | - |
| 20000 | NA | 0.93 | 0.53 | 0.62 | 0.69 | 0.74 | 0.75 | 0.08 | 0.00 | 0.03 | - |

## 表5：EnronQA基准子集的事实记忆。虽然目前，RAG是回忆事实信息的最佳方法，但训练LoRA适配器进行记忆可以匹配将所有事实放在上下文中的表现，这表明这是未来发展的一个有希望的

作为上下文集，并在真实的问题和答案对上进行测试。 对于长上下文，我们将所有我们试图记住的QA对（事实）放在上下文中，并将它们的答案放在Llama-3.1-8B-Instruct的上下文中。我们可以测试多达1,000个QA对，直到上下文长度满为止。对于RAG，我们在所有QA对上建立一个索引，并检索前100个（选择它是因为它是最适合长上下文的）与Llama-3.1-8B-Instruct上下文最相关的问答对。 我们使用ColBERTv2[[64](#_bookmark81)]作为检索器。最后，为了实现自动化，我们使用“Fictionary Unlearning的任务”论文中的设置来训练LoRA适配器，该论文测试了LoRA适配器上的unlearning [[48](#_bookmark65)]。 我们在所有事实上训练等级为{8，16，32，64，128，256，512，1024，2048}的LoRA适配器10个时期，速率为1×10 −4。 我们将alpha设置为秩的四倍，并使用0.05的dropout。 我们测试消融与哪一层适应，并发现做所有线性层的效果最好。所有设置均以LLama-3.1- 70 B-Instruct作为评判标准进行评估。

* 1. 结果

我们在表5中给出了该实验的结果[。](#_bookmark17)有趣的是，LoRA记忆可以在几乎­所有尺度上匹配长上下文性能，并继续超过阻止长上下文扩展的1000个QA对上限。事实上，对于许多LoRA适配器来说，性能只会开始降低大约20，000个记忆的事实，这表明仅仅LoRA参数就包含了惊人的容量。在所有尺度上，RAG都优于记忆和长上下文。这可能是由于任务的简单性（检索改写的QA对）以及当前RAG系统的强度。记忆化是一个相对未被充分研究的­现象（主要是用LLM来探索以试图*防止*记忆），因此，毫不奇怪，这还没有超过RAG。在未来，随着预训练和记忆方法的不断发展，通过LoRA适配器进行记忆可能会达到或超过RAG性能。

# 讨论

在这里，我们讨论了一些经验教训和宝贵的见解，研究人员在类似的问题。

**LLM自验证和优化管道可以是强大­的合成数据工具。** 我们的EnronQA基准测试由完全合成生成的问答对组成。过去

生成此类QA资源的过程需要大量的人力投入，或者需要从人们自然提问的平台（如Google [[41](#_bookmark58)]或Bing[[6](#_bookmark23)]）众包。相反，随着LLM功能的不断增长，我们能够将问题和答案的要求指定为可验证的单元测试。 这些问题需要“具体”、“客观­”、“有根据”和“高质量”。通过将这些检查作为单元测试编写，并端到端优化我们的系统以通过这些单元测试，我们能够在保持质量的同时综合生成大规模数据集。只有当问题完全成功地通过了所有四个单元测试­时，它们才成为我们的最终基准。 我们认为这是一种可扩展的模式：（1）将规范­写入单元测试，（2）优化管道（微调，提示等），（3）过滤基于单元测试的合成生成。这将是未来扩大数据收集工作的一种方式，这将严重依赖于单元测试本身的设计

## 通过微调或持续的预训练来实现自动化是未来检索的有趣方向当前

RAG的SOTA是检索此上下文，然后将其传递给LLM。然而，我们发现LLM能够记忆大量数据。例如，过去的RAG基准测试，如Natu-ralQuestions[[41](#_bookmark58)]和TriviaQA[[32](#_bookmark49)]都被LLM的参数知识所消耗。目前，这种参数化的­知识在很大程度上取决于互联网的组成，互联网是这些模型最大的训练数据来源。在未来，人们可以想象对私人文档进行持续的预训练，或者对记忆进行额外的微调。 在第6节中[，](#_bookmark16)我们展示了实现这一努力的一些初步步骤，并发现LoRA适配器可以在简化的设置中匹配长上下文来回忆事实知识。随着更多的工作继续预培训，我们

我希望EnronQA可以作为一种资源，用于测试这些方法，并在未来探索LLM记忆的局限性。

1. 结论

我们介绍了EnronQA，这是一个包含103，638封电子邮件的数据集，其中包含150个不同用户收件箱中的528，304个问答对。EnronQA能够更好地对RAG管道进行私人数据基准测试，并允许在实际数据上引入个性化检索设置的实验。 我们表明，EnronQA基准测试优于其他单跳检索基准测试，用于测量检索器和LLM的联合准确性。 我们­对现有的RAG管道进行基准测试，覆盖EnronQA上的一系列检索器、LLM和架构。最后，我们使用EnronQA探索权衡记忆和检索推理时，私人文件。我们将这个大型资源公开发布给社区，用于测试私人和个性化检索，并在持续的预训练中进行进一步研究，这是从大型语言模型的参数权重进行信息检索的潜在新前沿­

# 道德声明

EnronQA基准来自安然电子邮件语料库[[39](#_bookmark56)]，这是2003年西方能源市场调查的一部分并不是所有被泄露电子邮件的安然员工都犯有任何罪行，即使如此，我们仍然希望尊重安然电子邮件背后所有人的意愿，无论他们是否参与犯罪活动。

我们采取两个关键步骤来支持这些目标，尊重数据集背后的安然员工首先，我们使用2015年发布的数据集，其中根据要求从数据集中删除了几个人[17]。其次，我们应用过滤器从数据集中删除任何NSFW或有毒内容[（§3.1），](#_bookmark5)这可能是特别个人的。

除此之外，我们非常乐意支持任何受影响方的数据删除请求。如果出现任何此类删除请求，EnronQA数据集将持续维护和更新。安然电子邮件数据集已经在学术研究中使用了大约20年，我们希望支持继续道德地使用这一资源。

引用

1. CJ亚当斯，杰弗里索伦森，朱莉娅埃利奥特，卢卡斯迪克森，马克麦克唐纳，尼瑟姆，和威尔库基尔斯基。2017. 有毒评论分类挑战。[http：](https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge)

[//kaggle.com/competitions/jigsaw-toxic-comment-classification-挑战](https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge)

1. Vaibhav Adlakha，Shehzaad Dhuliawala，Kaheer Suleman，Harm de Vries，and Siva Reddy.2022年TopiOCQA：开放域会话问题分类与主题切换。*Transactions of the Association for Computational Linguistics*10（04 2022），468-483. <https://doi.org/10.1162/tacl_a_00471>arXiv：https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl\_a\_00471/2008126/tacl\_a\_00471.pdf
2. RavitejaAnantha，SvitlanaVakulenko，ZhuchengTu，ShayneLongpre，StephenPulman，andSrinivasChappidi. 2021年开放领域的提问通过问题重写进行对话。*计算语言学协会北美分会2021年会议论文集*：*人类语言技术*（2021）。
3. Simran Arora，Patrick Lewis，Angela Fan，Jacob Kahn，and Christopher Ré.2023年在基于检索的系统中对公共和私有数据的推理*计算语言学协会会刊*（2023）。[https://aclanthology.org/2023.tacl-1.51](https://aclanthology.org/2023.tacl-1.51/)
4. OrlandoAyala和PatriceBechard 2024年通过检索增强生成减少结构化输出中的幻觉。在计算语言学协会北美分会*2024年会议记录­­：人类语言技术（第6卷：行业跟踪）*，YiYang，AidaDavani，Avi Sil和Anoop Kumar（编辑）。 计算语言­学协会，墨西哥城，墨西哥，228-238。[https://doi.org/10.18653/v1/2024.naacl-www.example.com](https://doi.org/10.18653/v1/2024.naacl-industry.19)
5. Payal Bajaj，Daniel Campos，Nick Craswell，Li Deng，Jianfeng Gao，Xiaodong Liu，Rangan Majumder，Andrew McNamara，Bhaskar Mitra，Tri Nguyen，Mir Rosenberg，Xia Song，Alina Stoica，Saurabh Tivary，and Tong Wang.2018. MS MARCO：一个人工生成的机器阅读通信数据集。[arXiv：1611.09268](https://arxiv.org/abs/1611.09268)[cs.CL]<https://arxiv.org/abs/1611.09268>
6. Jon Ander Campos、Arantxa Otegi、Aitor Soroa、Jan Deriu、Mark Cieliebak和Eneko Agirre。2020. DoQA-通过对话式QA解决特定领域的常见问题。[arXiv：2005.01328](https://arxiv.org/abs/2005.01328)[cs.CL]<https://arxiv.org/abs/2005.01328>
7. Vittorio Castelli、Rishav Chakravarti、Saswati Dana、Anthony Ferritto、RaduFlorian、Martin Franz、Dinesh Garg、Dinesh Khandelwal、Scott McCarley、MikeMcCawley、MohamedNasr、LinPan、CezarPendus、JohnPitrelli、SaurabhPujar、Salim Roukos、Andrzej Sakrajda、Avirup Sil、Rosario Uceda-Sosa、Todd Ward和RongZhang。2019年。 TechQA数据集。[arXiv：1911.02984](https://arxiv.org/abs/1911.02984)[cs.CL][https：](https://arxiv.org/abs/1911.02984)

[//arxiv.org/abs/1911.02984](https://arxiv.org/abs/1911.02984)

1. Zhiyu Chen，Wenhu Chen，Charese Smiley，Sameena Shah，Iana Borova，DylanLangdon，Reema Moussa，Matt Beane，Ting-Hao Huang，Bryan Scholledge，andWilliam Yang Wang.2021. FinQA：金融数据的数值推理数据集。 在*2021年自然语言处理经验方法会议论文集中*，Marie-Francine Moens，Xuanjing Huang，Lucia Specia和ScottWen-tauYih（编辑）。 计算语言学协会，在线和蓬塔卡纳，多米尼加共和国，3697-3711。 [https://doi.org/10。18653/v1/2021.emnlp-main.300](https://doi.org/10.18653/v1/2021.emnlp-main.300)
2. Zhiyu Chen，Shiyang Li，Charese Smiley，Zhiqiang Ma，Sameena Shah，and William Yang Wang.2022. ConvFinQA：探索对话式金融问题推理中的数值推理在*2022年自然语言处理经验方法会议论文集*中，YoavGoldberg，Zornitsa Kozareva和Yue Zhang（编辑）。计算语言学协会，阿布扎比，阿拉伯联合酋长国，6279-6292。[http：](https://doi.org/10.18653/v1/2022.emnlp-main.421)

[//doi.org/10.18653/v1/2022.emnlp-www.example.com](https://doi.org/10.18653/v1/2022.emnlp-main.421)

1. Eunsol Choi，He He，Mohit Iyyer，Mark Yatskar，Wen-tau Yih，Yejin Choi，PercyLiang，and Luke Zettlemoyer.2018年QuAC：Question Questioning in Context。 在*2018年自然语言处理经验方法会议的会议记录中*，Ellen Hockoff，David Chiang，Julia Hockenmaier和Junichi Tsujii（编辑）。计算语言学协会，布鲁塞尔，比利时，2174-2184。<https://doi.org/10.18653/v1/D18-1241>
2. 马克·科利尔和乔瓦尼·比尔2019. 用于机器翻译的记忆增强神经网络 在*机器翻译峰会第十七届会议论文集：研究跟踪*，Mikel Forcada，Andy Way，Barry Haddow和Rico Sennrich（编辑）。 欧洲机器翻译协会，都柏林，爱尔兰，172-181。<https://aclanthology.org/W19-6617/>
3. 一起电脑2023年RedPajama*：一个用于训练大型语言模型的开放数据集­*。<https://github.com/togethercomputer/RedPajama-Data>
4. Vinay Deolalikar 2014. 距离还是覆盖？从企业文本集合中检索知识丰富的文档。在*第23届ACM­国际会议上的会议信息和知识管理*（上海，中国）*（CIKM'14）*。纽约计算机协会

美国纽约，1771-1774年。<https://doi.org/10.1145/2661829.2661865>

1. Abhimanyu Dubey，Abhinav Jauhri，Abhinav Pandey，Abhishek Kadian，Ah mad Al-Dahle，Aiesha Letman，Akhil Mathur，Alan Schelten，Amy Yang，Angela Fan，Anirudh Goyal，Anthony Hartshorn，Aobo Yang，Archi Mitra，Archie Sra-vankumar，Artem Korenev，Arthur Hinsvark，Arun Rao，Aston Zhang，Aurelien Rodriguez，Austen Gregerson，Ava Spataru，Baptiste Roziere，Bethany Biron，Binh Tang，Bobbie Chern，Charlotte Caucheteux，Chaya Nayak，Chloe Bi，Chris Marra，Chris McConnell，Christian Keller，Christophe Touret，Chunyang Wu，Corinne Wong，Cristian Canton Ferrer，Cyrus Nikolaeus，Damien Allonsius，Daniel Song，Danielle Pintz，Danny Livshits，David Esiobu，Dhruv Choudhary，Dhruv Maha-jan，Diego Garcia-Olano，Diego Perino，Dieuwke Hupkes，Egor Lakomkin，Ehab AlBadawy，Elina Lobanova，Emily Dinan，Eric Michael Smith，Filip Radenovic，Frank Zhang，Gabriel Synnaeve，Gabrielle Lee，Georgia Lewis Anderson，Graeme Nail，Gregoire Mialon，Guan Pang，Guillem Cucurell，Hailey Nguyen，Hannah Korevaar，Hu Xu，Hugo Touvron，Iliyan Zarov，Imanol Arrieta Ibarra，Isabel Kloumann，Ishan Misra，Ivan Evtimov，Jade Copet，Jaewon Lee，Jan Geffert，Jana Vranes，Jason Park，Jay Mahadeokar，Jeet Shah，Jelmer van der Linde，Jennifer Billock，Jenny Hong，Jenya Lee，Jeremy Fu，Jianfeng Chi，Jianyu Huang，Jiawen Liu，Jie Wang，Jiecao Yu，Joanna Bitton，Joe Spisak，Jongsoo Park，Joseph Rocca，Joshua Johnstun，Joshua Saxe，Junteng Jia，Kalyan Vasuden Alwala，Kartikeya Upasani，Kate Plawiak，Ke Li，Kenneth Heafield，Kevin Stone，Khalid El-Arini，Krithika Iyer，Kshitiz Malik，Kuenley Chiu，Kunal Bhalla，Lauren Rantala-Yeary，Laurens van der Maaten，Lawrence Chen，Liang Tan，Liz Jenkins，Louis Martin，Lovish Madaan，Lubo Malo，Lukas Blecher，Lukas Landzaat，Luke de Oliveira，Madeline Muzzi，Mahesh Pasupuleti，Mannat Singh，Manohar Paluri，Marcin­

Kardas，Mathew Oldham，Mathieu Rita，Maya Pavlova，Melanie Kambadur，Mike Lewis，Min Si，Mitesh Kumar Singh，Mona Hassan，Naman Goyal，Narjes Torabi，Nikolay Bashlykov，Nikolay Bogoychev，Niladri Chatterji，Olivier Duchenne，Onur Zelelebi，Patrick Alrassy，Pengchuan Zhang，Pengwei Li，Petar Vasic，Peter Weng，Prajjwal Bhargava，Pratik Dubal，Praveen Krishnan，Punit Singh Koura，Puxin Xu，Qing He，Qingxiao Dong，Ragavan Srinivasan，Raj Ganapathy，Ra mon Calderer，Ricardo Silveira Cabral，Robert Stojnic，Roberta Raileanu，Rohit Girdhar，Rohit Patel，Romain Sauvestre，Ronnie Polidoro，Roshan Sumbaly，Ross Taylor，Ruan Silva，Rui Hou，Rui Wang，Saghar Hosseini，Sahana Chennabasappa，Sanjay Singh，Sean Bell，Seohyun Sonia Kim，Sergey Eddom，Shaoliang Nie，Sharan Narang，Sharath Raparthy，Sheng Shen，Shengye Wan，Shruti Bhosale，Shun Zhang，Simon Vandenhende，Soumya Batra，Spencer Whitman，Sten Sootla，Stephane Collot，Suchin Gururangan，Sydney Borodinsky，Tamar Herman，Tara Fowler，Tarek Sheasha，Thomas Georgiou，Thomas Scialom，Tobias Speckbacher，Todor Mihaylov，Tong Xiao，Ujjwal Karn，Vedanuj Goswami，Vibhor Gupta，Vignesh Ramanathan，Viktor Kerkez，Vincent Gonguet，Vignie Do，Vish Vo-geti，Vladan Petrovic，Weiwei Chu，Wenhan Xiong，Wenyin Fu，Whitney Meers，Xavier Martinet，Xiaodong Wang，Xiaoqing Ellen Tan，Xinfeng Xie，Xuchao Jia，Xuwei Wang，Yaelle Goldschlag，Yashesh Gaur，Yasmine Babaei，Yi Wen，Yiwen Song，Yushen Zhang，Yue Li，Yunning Mao，Zacharie Delpierre Coudert，Zheng Yan，Zhengxing Chen，Zoe Papakipos，Aaditya Singh，Aaron Grattafiori，Abha Jain，Adam Kelsey，Adam Shajnfeld，Adithya Gangidi，Adolfo Victoria，Ahuva Goldstand，Ajay Menon，Ajay Sharma，Alex Boesenberg，Alex Vaughan，Alexei Baevski，Allie Feinstein，Amanda Kallet，Amit Sangani，Anam Yunus，An drei Lupu，Andres Alvarado，Andrew Caples，Andrew Gu，Andrew Ho，Andrew Poulton，Andrew Ryan，Ankit Ramchandani，Annie Franco，Aparajita Saraf，Arka-bandhu Chowdhury，Ashley Gabriel，Ashwin Bharambe，Assaf Eisenman，Azadeh Yazdan，Beau James，Ben Maurer，Benjamin Leonhardi，Bernie Huang，Beth Loyd，Beto De Paola，Bhargavi Paranjape，Bing Liu，Bo Wu，Boyu Ni，Braden Hancock，Bram Wasti，Brandon Spence，Brani Stojkovic，Brian Gamido，Britt Montalvo，Carl Parker，Carly Burton，Catalina Mejia，Changhan Wang，Changkyu Kim，Chao Zhou，Chester Hu，Ching-Hsiang Chu，Chris Cai，Chris Tindal，Christoph Feichtenhofer，Damon Civin，Dana Beaty，Daniel Kreymer，Daniel Li，Danny Wyatt，David Adkins，David Xu，Davide Testuggine，Delia David，Devi Parikh，Diana Liskovich，Didem Foss，Dingkang Wang，Duc Le，Dustin Holland，Edward Dowling，Eissa Jamil，Elaine Montgomery，Eleonora Presani，Emily Hahn，Emily Wood，Erik Brinkman，Esteban Arcaute，Evan Dunbar，Evan Smothers，Fei Sun，Felix Kreuk，Feng Tian，Firat Ozgenel，Francesco Caggioni，Francisco Guzmán，Frank Kanayet，Frank Seide，Gabriela Medina Florez，Gabriella Schwarz，Gada Badeer，Georgia Swee，Gil Halpern，Govind Thattai，Grant Herman，Grigory Sizov，Guangyi，Zhang，Guna Lakshminarayanan，Hamid Shojanazeri，Han Zou，Hannah Wang，Hanwen Zha，Haroun Habeeb，Harrison Rudolph，Helen Suk，Henry Aspegren，Hunter Goldman，Ibrahim Damlaj，Igor Molybog，Igor Tufanov，Irina-Elena Veliche，Itai Gat，Jake Weissman，James Geboski，James Kohli，Japhet Asher，Jean-Baptiste Gaya，Jeff Marcus，Jeff Tang，Jennifer Chan，Jenny Zhen，Jeremy Reizenstein，Jeremy Teboul，Jessica Zhong，Jian Jin，Jingyi Yang，Joe Cummings，Jon Carvill，Jon Shepard，Jonathan McPhie，Jonathan Torres，Josh Ginsburg，Junjie Wang，Kai Wu，Kam Hou U，Karan Saxena，Karthik Prasad，Kartikay Khandelwal，Katayoun Zand，Kathy Matosich，Kaushik Veeraraghavan，Kelly Michelena，Keqian Li，Kun Huang，Kunal Chawla，Kushal Lakhotia，Kyle Huang，Lailin Chen，Lakshya Garg，Lavender A，Leandro Silva，Lee Bell，Lei Zhang，Liangpeng Guo，Licheng Yu，Liron Moshkovich，Luca Wehrstedt，Madian Khabsa，Manav Avalani，Manish Bhatt，Maria Tsimpoukelli，Martynas Mankus，Matan Hasson，Matthew Lennie，Matthias Reso，Maxim Groshev，Maxim Nau-mov，Maya Lathi，Meghan Keneally，Michael L.­­Seltzer，Michal Valko，Michelle Restrepo，Mihir Patel，Mik Vyatskov，Mikayel Samvelyan，Mike Clark，Mike Macey，Mike Wang，Miquel Jubert Hermoso，Mo Metanat，Mohammad Rastegari，Munish Bansal，Nandhini Santhanam，Natascha Parks，Natasha White，Navyata Bawa，Nayan Singhal，Nick Egebo，Nicolas Bognier，Nikolay Pavlovich Laptev，Ning Dong，Ning Zhang，Norman Cheng，Oleg Bogguz，Olivia Hart，Omkar Salpekar，Ozlem Kalinli，Parkin Kent，Parth Parekh，Paul Saab，Pavan Balaji，Pedro Rittner，Philip Bontrager，Pierre Roux，Piotr Dollar，Polina Zvyagina，Prashant Ratanchandani，Pritish Yuvraj，Qian Liang，Rachad Alao，Rachel Rodriguez，Rafi Ayub，Raghotham Murthy，Raghu Nayani，Rahul Mitra，Raymond Li，Rebekkah Hogan，Robin Battey，Rocky Wang，Rohan Maheswari，Russ Howes，Ruty Rinott，Sai Jayesh Bondu，Samyak Datta、Sara Chugh、Sara Hunt、Sargun Dhillon、Sasha Sidorov、Satadru Pan、Saurabh Verma、Seiji Yamamoto、Sharadh Ramaswamy、Shaun Lindsay、Shaun Lindsay、Sheng Feng、Shenghao Lin、Shengxin Cindy Zha、Shiva Shankar、Shuqiang Zhang、Shuqiang Zhang，Sinong Wang，Sneha Agarwal，Soji Sajuyigbe，Soumith Chintala，Stephanie Max，Stephen Chen，Steve Kehoe，Steve Satterfield，Sudarshan Govindaprasad，Sumit Gupta，Sungmin Cho，Sunny Virk，Suraj Subramanian，Sy Choudhury，Sydney Goldman，Tal Remez，Tamar Glaser，Tamara Best，Thilo Kohler，Thomas Robinson，Tianhe Li，Tianjun Zhang，Tim Matthews，Timothy Chou，Tzook Shaked，Varun Vontimitta，Victoria Ajayi，Victoria Montanez，Vijai Mohan，Vinay Satish Kumar，Vishal Mangla，Vítor Albiero，Vlad Ionescu，Vlad Poenaru，Vlad Tiberiu Mihailescu，Vladimir Ivanov，Wei Li，Wenchen Wang，Wenwen Jiang，Wes Bouaziz，Will Constable，Xiaocheng Tang，Xiaofang Wang，Xiaojian Wu，Xiaolan Wang，Xide Xia，Xilun

Wu，Xinbo Gao，Yanjun Chen，Ye Hu，Ye Jia，Ye Qi，Yenda Li，Yilin Zhang，Ying Zhang，Yossi Adi，Youngjin Nam，Yu，Wang，Yuchen Hao，Yundi Qian，Yuzi He，Zach Rait，Zachary DeVito，Zef Rosnbrick，Zhaoduo Wen，Zhenyu Yang，and Zhiwei Zhao.2024年。The Llama 3 Herd of Models.[arXiv：2407.21783](https://arxiv.org/abs/2407.21783)[cs.AI]<https://arxiv.org/abs/2407.21783>

1. Matthew Dunn、Levent Sagun、Mike Higgins、V. Ugur Guney、Volkan Cirik和Kyunghyun Cho。2017. SearchQA：一个新的&问答数据集，增加了来自搜索引擎的上下文。[arXiv：1704.05179](https://arxiv.org/abs/1704.05179)[cs.CL]<https://arxiv.org/abs/1704.05179>
2. 安然公司和威廉W.科恩2015年。安然电子邮件数据集。[http：](https://www.loc.gov/item/2018487913/)

[//www.loc.gov/item/2018487913/](https://www.loc.gov/item/2018487913/)美国联邦能源监管委员会­代表团，威廉W.科恩，MLD，CMU，费城，宾夕法尼亚州。[软件，电子资源]。从国会图书馆检索。

1. Song Feng、Hui Wan、Chulaka Gunasekara、Siva Patel、Sachindra Joshi和Luis Lastras。2020年。doc2dial：一个面向目标的文档化对话数据集。在*2020年自然语言处理经验方法会议（EMNLP）上*，Bonnie Webber，Trevor Cohn，Yulan He和YangLiu（编辑）。 计算语言学协会，在线，8118-8128。<https://doi.org/10.18653/v1/2020.emnlp-main.652>
2. Leo Gao、Stella Biderman、Sid Black、Laurence Golding、Travis Hoppe、CharlesFoster、Jason Phang、Horace He、Anish Thite、Noa Nabeshima、Shawn Presser和ConnorLeahy。2020年。ThePile：An800GBDatasetofDiverseTextforLanguageModeling.*arXiv预印本arXiv：2101.00027*（2020）。
3. Yunfan Gao，Yun Xiong Gao，Xinyu Gao，Kangxiang Jia，Jinliu Pan，Yuxi Bi，Yi Dai，Jiawei Sun，Meng Wang，and Haofen Wang.2024年大型语言模型的检索增强生成：综述。[arXiv：2312.10997](https://arxiv.org/abs/2312.10997)[cs.CL]<https://arxiv.org/abs/2312.10997>
4. Samira Ghodratnama和MehrdadZakershahrak。2024.使LLM适应­高效，个性化的信息检索：方法和含义。 在*面向服务的计算-ICSOC 2023研讨会*，Flavia Monti，Pierluigi Plebani，Naouel Moha，Hye-young Paik，Johanna Barzen，Gowri Ramachandran，Devis Bianchini，Damian A. Tamburri，and Massimo Mecella（编辑）. Springer NatureSingapore，Singapore，17-26.
5. 大观研究所2024. 检索增强生成市场规模报告，2030年。<https://www.grandviewresearch.com/industry-analysis/retrieval-augmented-generation-rag-market-report>
6. 亚历克斯·格雷夫斯格雷格·韦恩和伊沃·丹尼赫尔卡2014年。神经图灵机[arXiv：1410.5401](https://arxiv.org/abs/1410.5401)[cs.NE]<https://arxiv.org/abs/1410.5401>
7. Kshitij Gupta，Benjamin Thérien，Adam Ibrahim，Mats L. Richter，Quentin Anthony，Eugene Belilovsky，Irina Rish，and Anasthée Lesort.2023年大型语言模型的持续预训练：如何（重新）预热模型？[arXiv：2308.04014](https://arxiv.org/abs/2308.04014)[cs.CL]<https://arxiv.org/abs/2308.04014>
8. Dan Hendrycks，Collin Burns，Anya Chen，and Spencer Ball.2021年CUAD：用于法律合同审查的专家注释NLP数据集。[arXiv：2103.06268](https://arxiv.org/abs/2103.06268)[cs.CL]<https://arxiv.org/abs/2103.06268>
9. 黄景，杨迪义，克里斯托弗·波茨。2024年揭秘大型语言模型中的逐字[arXiv：2407.17817](https://arxiv.org/abs/2407.17817)[cs.CL][https：](https://arxiv.org/abs/2407.17817)

[//arxiv.org/abs/2407.17817](https://arxiv.org/abs/2407.17817)

1. Yulong Hui，Yao Lu，and Huanchen Zhang.2024年UDA：真实世界文档分析中检索增强生成的基准套件[arXiv：2406.15187](https://arxiv.org/abs/2406.15187)[cs.AI]<https://arxiv.org/abs/2406.15187>
2. 无限流2024年RAGFlow：基于深度文档理解的开源RAG（检索增强生成­）引擎。 [https://github.com/](https://github.com/infiniflow/ragflow)[infiniflow/ragflow](https://github.com/infiniflow/ragflow)访问时间：2024-09-18。
3. Mohit Iyyer，Wen-Tau Yih，and Ming-Wei Chang.2017年。基于搜索的神经结构化学习的顺序问题分类。在*计算语言学协会第55届年会的会议记录（卷1：长论文）*。计算语言学协会，温哥华，加拿大，1821-1831。<https://doi.org/10.18653/v1/P17-1167>
4. 阿尔伯特·Q Jiang，AlexandreSablayrolles，AntoineRoux，ArthurMensch，BlancheSavary，ChrisBamford，DevendraSinghChaplot，DiegodelasCasas，EmmaBouHanna，Florian Bressand，Gianna Lengyel，Guillaume Bour，Guillaume Lample，Lélio Renard Lavaud，Lucile Saulnier，Marie-Anne Lachaux，Pierre Stock，SandeepSubramanian，SophiaYang，SzymonAntoniak，TenenLeScao，ThéophileGervet，Thibaut Lavril，Thomas Wang，Althée Lacroix，and William El Sayed.2024年。

混合专家。[arXiv：2401.04088](https://arxiv.org/abs/2401.04088)[cs.LG]<https://arxiv.org/abs/2401.04088>

1. Jiao Jin，Bhuwan Dhingra，Zhengping Liu，William Cohen，and Xinghua Lu. 2019年。PubMedQA：生物医学研究问题分类的数据集。在*2019年自然语言处理经验方法会议和第九届自然语言处理国际联合会议­（EMNLP-IJCNLP）的会议记录中*，Jiang，VincentNg和XiaojunWan（编辑）。计算语言学学会，中国香港，2567-2577。<https://doi.org/10.18653/v1/D19-1259>
2. MandarJoshi，EunsolChoi，DanielWeld，andLukeZettlemoyer. 2017年。triviaqa：一个用于阅读理解的大规模远程监督挑战数据集­。 *arXiv电子打印*，文章arXiv：1705.03551（2017），arXiv：1705.03551页。[1705.03551](https://arxiv.org/abs/1705.03551)
3. MandarJoshi，EunsolChoi，DanielWeld，andLukeZettlemoyer. 2017年。TriviaQA：一个用于阅读理解的大规模远程监督挑战数据集­在*第55届计算­语言学协会年会论文集（卷1：长论文）*，Regina Barzilay和Min-Yen Kan（编辑）。计算语言学协会，温哥华，加拿大，1601-1611。<https://doi.org/10.18653/v1/P17-1147>
4. 阿曼德·朱林，爱德华·格雷夫，彼得·博亚诺夫斯基，托马斯·米科洛夫。2016年。高效文本分类的技巧包。arXiv*预印本arXiv：1607.01759*（2016）。
5. EhsanKamalloo，ArefJafari，XinyuZhang，NandanThakur，andJimmyLin. 2023年HAGRID：一个人类LLM协作数据集，用于生成信息搜索与归因。*arXiv：2307.16883*（2023）。
6. Nikhil Kandpal，Haikang Deng，Adam Roberts，Eric Wallace，and Colin Raf fel.­2023. 大型语言模型努力学习长尾知识。[arXiv：2211.08411](https://arxiv.org/abs/2211.08411)[cs.CL]<https://arxiv.org/abs/2211.08411>
7. Zixuan Ke，Yijia Shao，Haowei Lin，Tatsuya Konishi，Gyuhak Kim，and Bing Liu.2023年语言模型的持续预训练*第十一届国际学习表征会议*。 [https://openreview.net/forum? id=m\_](https://openreview.net/forum?id=m_GDIItaI3o)[GDIItaI3o](https://openreview.net/forum?id=m_GDIItaI3o)
8. OmarKhattab，ArnavSinghvi，ParidhiMaheshwari，ZhiyuanZhang，KeshavSan­ thanam，Sri Vardhamanan，SaifulHaq，Ashutosh Sharma，Thomas T. Joshi，HannaMoazam，HeatherMiller，MateiZaharia，andChristopherPotts. 2023年DSPy：编译声明性语言模型调用到自我改进管道中。­*arXiv预印本arXiv：2310.03714*（2023）。
9. 布莱恩·克里姆特和杨一鸣。2004年 安然语料库：电子邮件分类研究的新数据集。 *欧洲机器学习会议*。SpringerBerlin Heidelberg，柏林，海德堡，217-226.
10. TomáKočisk，JonathanSchwarz，PhilBlunsom，ChrisDyer，KarlMoritzHer­ mann，GáborMelis，andEdwardGrefenstette. 2018年NarrativeQA阅读理解挑战 *Transactions of the Association for Computational* *Linguistics*6（2018），317-328. <https://doi.org/10.1162/tacl_a_00023>
11. TomKwiatkowski，JennimariaPalomaki，OliviaRedfield，MichaelCollins，AnkurParikh，ChrisAlberti，DanielleEpstein，IlliaPolosukhin，MatthewKelcey，JacobDevlin，WendonLee，KristinaN. Toutanova，LlionJones，Ming-WeiChang，AndrewDai，Jakob Uszkoreit，Quoc Le，and Slav Petrov.2019. 自然问题：问题推理研究的基准。*Transactions of the Association ofComputational Linguistics*（2019）.
12. Jure Leskovec，Anand Rajaraman，and Jeffrey David Ullman.2020年。*挖掘海量数据集*。北京：清华大学出版社.
13. Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、VladimirKarpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen tau Yih、Tim Rocktäschel、Sebastian Riedel和Douwe Kiela。2021. 知识密集型NLP任务的检索增强生成[arXiv：2005.11401](https://arxiv.org/abs/2005.11401)[cs.CL]<https://arxiv.org/abs/2005.11401>
14. Jimmy Lin，Xuoguang Ma，Sheng-Chieh Lin，Jheng-Hong Yang，Ronak Pradeep，andRodrigoNogueira. 2021年Pyserini：一个用于­稀疏和密集表示的可再现信息检索研究的Python工具包。第*44届国际ACM SIGIR信息检索研究与开发会议­**（SIGIR 2021）*2356-2362。
15. SijiaLiu，YuanshunYao，JinghanJia，StephenCasper，NathalieBaracaldo，PeterHase，Yuguang Yao，Chris Yuhao Liu，Xiaojun Xu，Hang Li，Kush R.Varshney，MohitBansal，SanmiKoyejo，andYangLiu. 2024年重新思考大型语言模型的机器非学习。arXiv[：2402.08787](https://arxiv.org/abs/2402.08787)[cs.LG][https://arxiv.org/abs/2402。](https://arxiv.org/abs/2402.08787) [08787](https://arxiv.org/abs/2402.08787)
16. Yougang Lyu，Lingyong Yan，Shuaiqiang Wang，Haibo Shi，Dawei Yin，Pengjie Ren，Zhumin Chen，Maarten de Rijke，and Zhaochun Ren.2024. KnowTun-ing：大型语言模型的知识感知微调。在*2024年自然语言处理经验方法会议上*，Yaser Al-Onaizan，Mohit Bansal和Yun-Nung Chen（编辑）。­计算语言学协会，迈阿密，佛罗里达州，美国，14535-14556。[http：](https://doi.org/10.18653/v1/2024.emnlp-main.805)

[//doi.org/10.18653/v1/2024.emnlp-www.example.com](https://doi.org/10.18653/v1/2024.emnlp-main.805)

1. 马新北，龚业云，何鹏程，赵海，段南。2023年检索增强的大型语言模型中的查询重写。在*2023年*自然语言处理经验方法会议*的会议记录中*，HoudaBouamor，Juan Pino和Kalika Bali（编辑）。计算语言学协会，新加坡，5303-5315。[https://doi.org/10.18653/v1/2023.emnlp-www.example.com](https://doi.org/10.18653/v1/2023.emnlp-main.322)
2. PratyushMaini，ZhiliFeng，AviSchwarzschild，ZacharyChaseLipton，andJZicoKolter. 2024. TOFU：法学硕士的虚拟遗忘任务。第*一届**语言建模会议*。 [https://openreview.net/forum? id=B41hNBoWLo](https://openreview.net/forum?id=B41hNBoWLo)
3. ChaitanyaMalaviya，SubinLee，SihaoChen，ElizabethSieber，MarkYatskar，andDanRoth. 2024年ExpertQA：专家策划的问题和归因的答案。[arXiv：2309.07852](https://arxiv.org/abs/2309.07852)[cs.CL]<https://arxiv.org/abs/2309.07852>
4. 卢克·梅里克2024年。 嵌入和聚类数据可以改善对比预训练。[arXiv：2407.18887](https://arxiv.org/abs/2407.18887)[cs.LG]<https://arxiv.org/abs/2407.18887>
5. Timo Möller，Anthony Reina，Raghavan Jayakumar，and Malte Pietsch.2020年。COVID-QA：COVID-19的问答数据集。在*ACL 2020的第一届COVID-19NLP研讨会上*，KarinVerspoor，KevinBretonnelCohen，MarkDredze，EmilioFerrara，JonathanMay，RobertMunro，CecileParis，

和拜伦·华莱士（编辑）。计算语言学协会，在线。<https://aclanthology.org/2020.nlpcovid19-acl.18>

1. Chenghao Mou，Chris Ha，Kenneth Enevoldsen，and Peiyuan Liu. 2023. ChenghaoMou*/text-dedup：参考快照*。 [https://doi.org/10.5281/zenodo。8364980](https://doi.org/10.5281/zenodo.8364980)
2. KaiNakamura，SharonLevy，Yi-LinTuan，WenhuChen，andWilliamYangWang. 2022年HybridDialogue：基于表格和文本数据的信息搜索对话数据集。在计算语言学协会的*调查结果：ACL 2022*，Smaranda Muresan，Preslav Nakov和Aline Villavicencio（编辑）。 计算语言学协会，都柏林，爱尔兰，481-492。[http：](https://doi.org/10.18653/v1/2022.findings-acl.41)

[//doi.org/10.18653/v1/2022.findings-www.example.com](https://doi.org/10.18653/v1/2022.findings-acl.41)

1. Abhilash Nandy、Soumya Sharma、Shubham Maddhashiya、Kapil Sachdeva、Pawan Goyal和NIloy Ganguly。2021年 电子设备上的问答：一个新的基准数据集和基于多任务学习的问答框架­。在计算语言学协会的*调查结果：EMNLP2021*，Marie-Francine Moens，Xuanjing Huang，Lucia Specia和Scott Wen-tau Yih（编辑）。计算语言学协会，蓬塔卡纳，多米尼加共和­国，4600-4609。<https://doi.org/10.18653/v1/2021.findings-emnlp.392>
2. Krista Opsahl-Ong，Michael J Ryan，Josh Purtell，David Broman，Christo pher Potts，Matei Zaharia，and Omar Khattab. 2024. 多阶段语言模型程序的优化指令和演示。[arXiv：2406.11695](https://arxiv.org/abs/2406.11695)[cs.CL]<https://arxiv.org/abs/2406.11695>
3. FabioPetroni、Aleksandra Piktus、Angela Fan、Patrick Lewis、Majid Yazdani、Nicola De Cao、James Thorne、Yacine Jernite、Vladimir Karpukhin、Jean Mail-lard、VassilisPlachouras、TimRocktäschel和SebastianRiedel。2021年KILT：知识密集型语言任务的基准。 在*计算语言学协会北美分会2021年会议论文集：人类语言技术*，Kristina Toutanova，Anna Rumshisky，Luke Zettlemoyer，DilekHakkani-Tur，Iz Beltagy，Steven Bethard，Ryan Cotterell，Tanmoy Chakraborty和Yichao Zhou（编辑）。计算语言学协会，在线，2523-2544。<https://doi.org/10.18653/v1/2021.naacl-main.200>
4. FabioPetroni、TimRocktäschel、SebastianRiedel、PatrickLewis、AntonBakhtin、YuxiangWu和AlexanderMiller。2019年。 作为知识基础的语言模型？ 在*2019年自然语言处理经验方法会议和第九届自然语言处理国际联合会议­（EMNLP-IJCNLP）的会议记录中*，Jiang，VincentNg和XiaojunWan（编辑）。计算语言学学会，香港，中国，2463-2473。<https://doi.org/10.18653/v1/D19-1250>
5. 法比奥·皮内利，加布里埃尔·托洛梅，乔瓦尼·特拉波里尼。2023. FLIRT：联邦学习信息检索。第*46届国际ACMSIGIR信息检索研究与发展会议*（台北，台湾）*（SIGIR '23）*。 计算机协会，纽约，纽约，美国，3472-3475。<https://doi.org/10.1145/3539618.3591926>
6. 杰克·W Rae，Sebastian Borgeaud，Trevor Cai，Katie Millican，Jordan Hoffmann，Francis Song，John Aslanides，Sarah Henderson，Roman Ring，Susannah Young，Eliza Rutherford，Tom Hennigan，Jacob Menick，Albin Cassirer，Richard Powell，George van den Driessche，Lisa Anne Hendricks，Maribeth Rauh，Po-Sen Huang，Amelia Glaese，Johannes Welbl，Sumanth Dathathri，Saffron Huang，Jonathan Ue-sato，John Mellor，Irina Higgins，Antonia Creswell，Nat McAleese，Amy Wu，Erich Elsen，Siddhant Jayakumar，Elena Buchatskaya，David Budden，Esme Sutherland，Karen Simonyan，Michela Paganini，Laurent Sifre，Lena Martens，Xiang Lorraine Li，Adhiguna Kuncoro，Aida Nematzadeh，Elena Gribovskaya，Domenic Donato，Angeliki Lazaridou，Arthur Mensch，Jean-Baptiste Lespiau，Maria Tsimpoukelli，Nikolai Grigorev，Doug Fritz，Thibault Sottiaux，Mantas Pajarskas，Toby Pohlen，Zhitao Gong，Daniel Toyama，Cyprien de Masson dAutume，Yujia Li，Tayfun Terzi，Vladimir Mikulik，Igor Babuschkin，Aidan Clark，Diego de Las Casas，Au relia Guy，Chris Jones，James Bradbury，Matthew Johnson，Blake Hechtman，Laura Weidinger，Iason Gabriel，William Isaac，埃德·洛克哈特、西蒙·奥辛德罗、劳拉·里梅尔、克里斯·戴尔、奥里奥尔·维纳尔斯、卡里姆·阿尤布、杰夫·斯坦韦、洛林·贝内特、德米斯·哈萨比斯、科雷·卡武库奥卢和杰弗里·欧文。2022年。缩放­语言模型：方法，分析训练地鼠的见解。[arXiv：2112.11446](https://arxiv.org/abs/2112.11446)[cs.CL]<https://arxiv.org/abs/2112.11446>
7. Pranav Rajpurkar，Robin Jia，Percy Liang. 2018. 知道你不知道的：SQuAD无法回答的问题。[arXiv：1806.03822](https://arxiv.org/abs/1806.03822)[cs.CL][https：](https://arxiv.org/abs/1806.03822)

[//arxiv.org/abs/1806.03822](https://arxiv.org/abs/1806.03822)

1. Siva Reddy，Danqi Chen和Christopher D.曼宁2019. CoQA：一个传统的提问挑战。 *Transactions of the Association for* *Computational Linguistics* 7（2019），249-266. <https://doi.org/10.1162/tacl_a_00266>
2. 亚当·罗伯茨科林·拉菲尔和诺姆·沙泽尔2020年。你能把多少知识打包到语言模型的参数中？在*2020年自然语言处理经验方法会议（EMNLP）上*，Bonnie Webber，Trevor Cohn，Yulan He和Yang Liu（编辑）。 计算语言学协会，在线，5418-5426。 [https://doi.org/10.18653/v1/2020。](https://doi.org/10.18653/v1/2020.emnlp-main.437)emnlp—main[.437](https://doi.org/10.18653/v1/2020.emnlp-main.437)
3. Mobashir Sadat，Zhengyu Zhou，Lukas Lange，Jun Araki，Arsalan Gundroo，Bingqing Wang，Rakesh Menon，Md Parvez，and Zhe Feng. 2023. DelucionQA：检测特定领域问题分类中的幻觉。计算*语言学协会（Association for Computational Linguistics：EMNLP 2023*）协会

计算语言学，新加坡，822-835。<https://doi.org/10.18653/v1/2023.findings-emnlp.59>

1. Keshav Santhanam，Omar Khattab，Jon Saad-Falcon，Christopher Potts和MateiZaharia。2022年ColBERTv2：EffectiveandEfficientRetrievalviaLightweightLateInteraction. 在计算语言学协会*北美分会2022年会议的会议记录中：人类语言技术*，Marine Carpuat，Marie-Catherine de Marneffe和Ivan Vladimir Meza Ruiz（编辑）。 计算语言学协会，西雅图，美国，3715-3734。<https://doi.org/10.18653/v1/2022.naacl-main.272>
2. 米拉德·肖库希和罗斯。2011.联邦搜索。 *Foundationsand Trends®in**Information Retrieval*5，1（2011），1-102. <https://doi.org/10.1561/1500000010>
3. Kurt Shuster，Spencer Poff，Moya Chen，Douwe Kiela和Jason Weston。2021年提取增强减少对话中的幻觉 *arXiv预印本**arXiv：2104.07567*（2021）。<https://arxiv.org/abs/2104.07567>
4. LucaSoldaini，Rodney Kinney，Akshita Bhagia，Dustin Schwenk，David Atkin­son，RussellAuthur，BenBogin，KhyathiBagu，JenniferDumas，YanaiElazar，ValentinHofmann，AnanyaHarshJha，SachinKumar，LiLucy，XinxiLyu，NathanLambert，IanMagnusson，JacobMorrison，NiklasMuennighoff，AakankshaNaik，Crystal Nam，Matthew E.作者：Peter，Abhilasha Ravichander，Kyle Richardson，Ze-jiang Shen，Emma Strubell，Nishant Subramani，Oyvind Tafjord，Pete Walsh，LukeZettlemoyer，NoahA. 史密斯，汉纳内·哈吉希尔齐，伊兹·贝尔塔吉，德克·格伦-埃维尔德，杰西·道奇和凯尔·罗。2024年Dolma：An Open Corpus of Three Tril Lion Tokens for Language Model Pretraining Research（英语：Dolma：An Open Corpus of Three Tril Lion Tokens for Language Model Pretraining Research）*ArXiv预印本*（2024）。<https://arxiv.org/abs/2402.00159>
5. Viju Sudhi、Sinchana Ramakanth Bhat、Max Rudat和Roman Teucher。2024. RAG-Ex：一个解释检索增强生成的通用框架。在*第47届国际ACM SIGIR会议的研究和发展信息检索*（华盛顿特区，美国）*（SIGIR'24）*。计算机协会，纽约，纽约，美国，2776-2780。<https://doi.org/10.1145/3626772.3657660>
6. 凯瑟琳·田，埃里克·米切尔，姚华秀，克里斯托弗·曼宁和切尔西·芬恩。2023。为真实性微调语言模型。在*NeurIPS 2023年关于指令调整和指令遵循的研讨会上*。 [https://openreview.net/forum? id=kEK08VdSO5](https://openreview.net/forum?id=kEK08VdSO5)
7. AdamTrischler，Tong Wang，Xingdi Yuan，Justin Harris，Alessandro Sordoni，PhilipBachman，andKaheerSuleman. 2017年。NewsQA：一个机器压缩­数据集。In *Proceedingsofthe2ndWorkshoponRepresentationLearningforNLP*，Phil Blunsom，Antoine Bordes，Kyunhyun Cho，Shay Cohen，Chris Dyer，Edward Grefenstette，Karl Moritz Hermann，Laura Rimell，Jason Weston，and ScottYih（Eds.）. 计算语言学协会，温哥华，加拿大，191-200。<https://doi.org/10.18653/v1/W17-2623>
8. ShuaiWang，EkaterinaKhramtsova，ShengyaoZhuang，andGuidoZuccon. 2024年。FeB4RAG：在检索增强生成的背景下评估联邦搜索 在*第47届国际ACM SIGIR信息检索研究与发展会议*（美国华盛顿特区）*（SIGIR'24）的会议记录中*。计算机协会，纽约，纽约，美国，763-773。[http：](https://doi.org/10.1145/3626772.3657853)

[//doi.org/10.1145/3626772.3657853](https://doi.org/10.1145/3626772.3657853)

1. Zijie J. Wang和Duen Horng Chau。2024. MeMemo：用于私人和个性化文本生成的设备上检索增强在*第47届国际ACM SIGIR信息检索研究与发展会议*（美国华盛顿特区）*（SIGIR '24）*的会议记录中。计算机协会，纽约，纽约，美国，2765-2770。[https://doi.org/10.1145/www.example.com](https://doi.org/10.1145/3626772.3657662)
2. HuiWei，ShenghuaHe，TianXia，AndyWong，JingyangLin，andMeiHan. 2024年。LLM对齐任务中LLM-as-a-Judge的系统评估：可解释的错误和多样的提示模板。[arXiv：2408.13006](https://arxiv.org/abs/2408.13006) [cs.CL][https：//arxiv.](https://arxiv.org/abs/2408.13006) org[/abs/2408.13006](https://arxiv.org/abs/2408.13006)
3. ZeqiuWu，Ryu Parish，Hao Cheng，Sewon Min，Prithviraj Ammanabrolu，Mari Ostendorf，andHannanehHajishirzi. 2023年InSCIt：信息寻求­与混合主动互动的对话。Transactions*of the Association for­* *Computational Linguistics*11（2023），453-468. <https://doi.org/10.1162/tacl_a_00559>
4. Sirui Xia，Xintao Wang，Jiaqing Liang，Yifei Zhang，Weikang Zhou，Jiaji Deng，Fei Yu，and Yanghua Xiao.2024. Ground Every Sentence：­ImprovingRetrieval-AugmentedLLMwithInterleavedReference-ClaimGeneration. [arXiv：2407.01796](https://arxiv.org/abs/2407.01796)[cs.CL]<https://arxiv.org/abs/2407.01796>
5. 杨梓潼，Neil Band，李双平，Emmanuel Candes和Tatsunori Hashimoto。2025年合成人继续训练。 在*第十三届­国际学习表征会议上*。 [https://openreview.net/forum?](https://openreview.net/forum?id=07yvxWDSla) ID[=07yvxWDSla](https://openreview.net/forum?id=07yvxWDSla)
6. Zhilin Yang，Peng Qi，Saizheng Zhang，Yoonge Bengio，William Cohen，Ruslan Salakhutdinov，and Christopher D.曼宁2018年HotpotQA：一个用于多样化，可解释的多跳问题分类的数据集。在*2018年自然语言处理经验方法会议的会议记录*中，Ellen Hockoff，David Chiang，Julia Hockenmaier和Junichi Tsujii（编辑）。计算语言学协会，布鲁塞尔，比利时，2369-2380。 [https://doi.org/10。](https://doi.org/10.18653/v1/D18-1259)18653[/v1/D18-1259](https://doi.org/10.18653/v1/D18-1259)
7. Shenglai Zeng，Jiankun Zhang，Pengfei He，Jie Ren，Tianqi Zheng，HanqingLu，Han Xu，Hui Liu，Yue Xing，and Jiliang Tang. 2024. 通过纯合成数据缓解检索­增强生成（RAG）中的隐私问题。[arXiv：2406.14773](https://arxiv.org/abs/2406.14773)[cs.CR]<https://arxiv.org/abs/2406.14773>
8. ShenglaiZeng，JiankunZhang，PengfeiHe，YueXing，YidingLiu，HanXu，JieRen，Shuaiqiang Wang，Dawei Yin，Yi Chang，and Jiliang Tang.2024年。 好与坏：探索检索增强生成（RAG）中的隐私问题。[arXiv：2402.16893](https://arxiv.org/abs/2402.16893)[cs.CR]<https://arxiv.org/abs/2402.16893>
9. Saber Zerhoudi和Michael Granitzer2024年。PersonaRAG：使用以用户为中心的代理增强检索增强生成系统。[arXiv：2407.09394](https://arxiv.org/abs/2407.09394)[cs.IR]<https://arxiv.org/abs/2407.09394>
10. Fengbin Zhu，Wenqiang Lei，Youcheng Huang，Chao Wang，Shuo Zhang，Jiancheng Lv，Fuli Feng，and Tat-Seng Chua.2021. TAT-QA：一个问题，一个关于金融中表格和文本内容混合的标准 在*Proceedings of the 59th Annual Meeting of the Association for Computational Linguisticsandthe11thInternationalJointConferenceonNaturalLanguageProcessing­（Volume1：Long Papers）*，Chengqing Zong，FeiXia，WenjieLi，andRobertoNavigli（Eds.）.计算语言学协会，在线，3277-3287。<https://doi.org/10.18653/v1/2021.acl-long.254>

# 与其他QA和RAG基准的比较

表[6](#_bookmark100)包含了与其他流行的QA和RAG基准的比较。EnronQA使用私人电子邮件覆盖了未开发的私人知识领域。与其他资源相比，它有相当或更多的­文件，同时涵盖了更多的问题。每个文档有多个问题将有助于训练记忆文档中的事实信息，并使研究能够微调和优化RAG管道，而不仅仅是作为诊断基准。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 基准 | 语料库大小 | QA配对/旋转 | 域 | 源 |
| ConcurrentQA[[4]](#_bookmark21) | 5.2百万+47千 | 18.4k | 一般+私人知识 | 维基百科+电子邮件 |
| ConvFinQA[[10]](#_bookmark27)  [[61]](#_bookmark78)第六十一话  [[51]](#_bookmark68)第五十一话 | 2k8.4k  147 | 14k127k  2k | 金融  一般知识学术研究 | 财务报告  文学，学术界，新闻，维基百科，Reddit，考试研究论文 |
| CUAD [25] | 510 | 13k | 法律 | 法律合同 |
| [[63]](#_bookmark80)第六十三话 | 1 | 2k | 客户支持 | 吉普车手册 |
| [[第18话]](#_bookmark35) | 458 | 25.7k | 政府 | 政府网站 |
| DoQA[[7]](#_bookmark24) | 2.4k | 10.9k | 烹饪，旅游，电影 | Stack Exchange |
| 电子手册[[54]](#_bookmark71) | 308k | 3.3k | 客户支持 | TVManual |
| [[49]](#_bookmark66)第四十九话 | – | 2.2k | 专家知识 | 谷歌搜索 |
| FinQA[[9]](#_bookmark26) | 2.8k | 8.3k | 金融 | 财务报告 |
| 哈格里克[[35]](#_bookmark52) | 32.8M | 2.6k | 一般知识 | 维基百科 |
| [[第77话]](#_bookmark94) | 5.2M | 112.8k | 一般知识 | 维基百科 |
| [[53]](#_bookmark70)第五十三话 | 2.9k | 22.5k | 一般知识 | 维基百科 |
| INSCIT[[74]](#_bookmark91) | 6.6M | 4.7k | 一般知识 | 维基百科 |
| 马可女士[[6]](#_bookmark23)  [[40]](#_bookmark57)第四十话 | 3.6M  1.6k | 1.01M  46.8k | 一般知识  电影，文学 | 古腾堡网页计划  +IMSDB |

金融、学术、知识基础

金融，研究论文，维基百科

EnronQA（Ours）103.6k 528.3k 私人知识电子邮件

表6：基于文档的QA基准比较。 EnronQA涵盖了与许多流行的QA基准相当或更大的语料库规模，同时拥有更多的QA对，可以进行训练，优化和文档记忆探索。此外，EnronQA使用电子邮件跨越了未开发的私人文档领域

# 语言模型库

* 1. 初始QA生成提示

在这里，我们包括用于QA生成的未优化提示和DSPyMIPROv2[[55](#_bookmark72)]优化提示，包括重写的指令和四个自举的几个镜头示例。 此提示是通过基于电子邮件和先前问题的区别创建初始问题来播种问题细化过程。我们使用30封电子邮件和20封验证电子邮件的训练集进行优化。我们运行MIPROv2进行20次迭代（批处理），并生成10条候选指令进行搜索。我们使用Llama3.1 70b指令[[15](#_bookmark32)]作为提示生成器模型。

初始QA生成提示（DSPy优化）（第1部分）

想象一下，你是一个高风险的商业顾问，你必须提出一个问题来测试读者对这封电子邮件的理解，以便提供完成一笔数百万美元交易所需的关键信息。你应该把这个问题

\* 完全真实 \*，即不提供意见，测试读者自己对问题的想法，或要求电子邮件的结论。不要暗示你感到困惑。此外，这个问题不能用主观判断来回答。但是，可以使用复合疑问句，例如。“什么是X，什么是Y？”快速而中肯地写出这样一个问题将意味着获得一笔可观的奖金。给定一封特定的电子邮件和一组我已经问过的关于这封电子邮件的问题，写一个新的问题以确保成功。

-

遵循以下格式。

电子邮件：我们想测试读者对电子邮件的理解的电子邮件。

优先问题：我已经就这封电子邮件询问过的优先问题推理：让我们一步一步思考，以便${produce the new\_question}。我们... 新问题：我们想问的关于电子邮件的全新问题

-

电邮：主题：午餐

邮件：[buylow@houston.rr.com](mailto:buylow@houston.rr.com)收件人：['[dfarmer@enron.com']](mailto:dfarmer@enron.com)文件：farmer-d/personal/292.

树胡子的

我被迷住了。指定见面的时间和地点替我祝阿美好运。

谁能想到他这么快就回来了。现在我们必须与他们，K州，和会议冠军（如果我们生存）。如果我们不放松彼此，我看不出12强是第一名

你想笑一笑--我女儿是一个大NU球迷，已经去看了几场贝勒的比赛。他说这和林肯不太一样;但是

期望不高。

午饭的时候告诉我我会去的。肯

=====================================

提问：无;问任何问题

推理：让我们一步一步地思考，以产生新的问题。我们可以先看看电子邮件的内容。发件人正在讨论一个午餐会，并提到了一些运动队，包括K-State和K-State。他们还提到，他们的女儿是一个球迷的NU，并参加了贝勒游戏。 我们可以问一个问题，测试读者对发件人家庭的理解。

新问题：发送者的女儿是哪支球队的球迷？

-

初始QA生成提示（DSPy优化）（第2部分）

电子邮件：主题：回复：fritzbrinkman简历邮箱：[mike. enron.com](mailto:mike.mcconnell@enron.com)收件人：[robert. jones@enron.com ']](mailto:robert.jones@enron.com)文件：mcconnell-m/all\_documents/442。

=====================================

罗拔，

谢谢，我已经和汤姆·怀特谈过了，我会把他的简历转发给你考虑.我也会和斯坦·霍顿谈谈。

迈克

=====================================

先前的问题：1.麦克·麦康奈尔说他会把简历转发给谁考虑，他还打算和谁谈？

推理：让我们一步一步地思考，以产生新的问题。我们需要分析电子邮件和前面的问题，以确定尚未涵盖的电子邮件的新方面。前一个问题集中在迈克麦康奈尔将采取的行动与简历。一个新的问题可以集中在电子邮件的上下文，如发件人和收件人之间的关系。

新问题：根据关于弗里茨布林克曼简历的电子邮件的语气和内容，迈克·麦康奈尔和罗伯特·琼斯之间的关系是什么

-

电子邮件：主题：Asm Dem Plan电子邮件：[jeff. enron.com](mailto:jeff.dasovich@enron.com)

收件人：[[[“richard. enron.com”]](mailto:richard.shapiro@enron.com)文件：dasovich-j/sent/4786。

=====================================

这是我今天从说客那里收到的文件这是我昨天在内部分发的同一份文件。

最好的，杰夫

--由Jeff Dasovich/NA/Enron于05/31/2001 05：26 PM转发--Delaney Hunter”[dhunter@s-k-w.com](mailto:dhunter@s-k-w.com)

2001年5月30日下午 12时29分

请回复dhunter

收件人：Jeff Dasovich（电子邮件）[jdasovic@enron.com](mailto:jdasovic@enron.com)抄送：

主题：Asm Dem计划

一个小东西，我们得到了我们的手...

- 装配演示计划. doc

=====================================

先前的问题：1.杰夫·达索维奇从说客那里收到的文件的名称是什么，最初是谁转发给他的？

推理：让我们一步一步地思考，以产生新的问题。 我们可以分析电子邮件，以确定先前问题未涵盖的关键要素。前一个问题集中在文件的名称和转发者。 我们可以在电子邮件中查找可用于创建新问题的其他具体细节。 一个突出的细节是电子邮件由JeffDasovich转发的日期和时间。 我们可以创建一个问题来测试读者对这个具体细节的理解。新问题：杰夫·达索维奇是在什么日期和时间将说客的文件转发给理查德·夏皮罗的？

初始QA生成提示（DSPy优化）（第3部分）

-

电子邮件：主题：服务交易

收件人：[holden. enron.com](mailto:holden.salisbury@enron.com)收件人：[portland. shift@enron.com ']](mailto:portland.shift@enron.com)文件：guzman-m/all\_documents/1285。

=====================================

如果您正在为Harbor、LV Cogen、Willamette或Saguaro进行交易，请复制

服务表上的交易模板工作表中的交易。这些交易经过编码以显示我们的利润。你可以随意更改任何单元格，

以蓝色突出显示的买入价格和损益。这些单元格具有不应更改的公式。请确保您输入正确的基础价格以及。

谢谢，霍顿

=====================================

提问：无;问任何问题

推理：让我们一步一步地思考，以产生新的问题。我们需要确定电子邮件的主题，即关于为特定发电厂进行交易的主题。我们还需要考虑发送者试图传达什么信息，这是关于使用模板和遵循某些指导方针。我们可以问一个问题，测试读者对电子邮件中提到的具体指导方针的理解

新问题：根据Holden关于在服务表中输入交易的电子邮件，交易模板中的哪两个单元格不应由用户更改？

-

电子邮件：{**email**}

推理：让我们一步一步地思考，以便

# 电子邮件选择提示

此提示用于衡量问题的具体性。如果Llama3.1 70b能够从10封电子邮件的列表中选择对应于问题的正确电子邮件，则我们认为该问题是具体的。

# 针对特异性提示的QA细化

在这里，我们包括未优化的提示，用于QA细化，使问题更具体，以及DSPy MIPROv2 [[55](#_bookmark72)]优化的提示，包括重写的指令和一个引导的fewshot示例。 此提示用于重写问题，以便它们更加具体，并且不会意外引用多个不同的电子邮件（或由多个不同的电子邮件回答）。 这在§B.1中描述的相同端到端优化中进行优化[。](#_bookmark101)

针对特异性的QA优化（DSPy优化）（第1部分）

你是一个高风险的调查员，任务是揭露一系列神秘电子邮件背后的真相。 您对特定的电子邮件有疑问，但您需要对其进行优化，以确保它特定于该电子邮件，而不适用于其他类似的电子邮件。整个调查的成败就取决于你是否有能力提出准确的问题。 给定电子邮件和一组不相关但相似的电子邮件，细化问题，使其更具体地针对所讨论的确切电子邮件，并在必要时添加电子邮件中的细节。这个问题必须用一句话来回答整个调查都指望你了仔细推敲问题

-

遵循以下格式。

电子邮件：我们要优化问题的电子邮件问题：我们要优化的问题

其他电子邮件：我们不希望问题涉及的其他电子邮件。确保优化后的问题包含不适用于这些电子邮件的详细信息

推理：让我们一步一步地思考，以便${产生refined\_question}。我们... 精炼问题：更具体到有问题的电子邮件的精炼问题

-

电子邮件：主题：Asm Dem Plan电子邮件：[jeff. enron.com](mailto:jeff.dasovich@enron.com)

收件人：[[[“richard. enron.com”]](mailto:richard.shapiro@enron.com)文件：dasovich-j/sent/4786。

=====================================

这是我今天从说客那里收到的文件这是我昨天在内部分发的同一份文件。

最好的，杰夫

--由Jeff Dasovich/NA/Enron于05/31/2001 05：26 PM转发--Delaney Hunter[dhunter@s-k-w.com](mailto:dhunter@s-k-w.com)

2001年5月30日下午 12时29分

请回复dhunter

收件人：Jeff Dasovich（电子邮件）[jdasovic@enron.com](mailto:jdasovic@enron.com)抄送：

主题：Asm Dem计划

一个小东西，我们得到了我们的手...

- 装配演示计划. doc

=====================================

问：Jeff Dasovich在什么日期和时间将电子邮件转发给RichardShapiro？

其他电子邮件：\u 00 ab 1\ u00bb.test-原始信息来自：dasovich，jeff发送时间：2001年11月21日，星期三，12：37

下午传译：杰夫达索维奇主题：测试

\u00ab2\ u00bb. - ---原始信息来自：dasovich，jeff发送时间：2001年10月6日星期六下午2：44，发送至：“jdasovich@

sbcglobal。net'主题：

\u00ab3\u00bb.这是我今天从说客那里收到的文件。这是我昨天在内部分发的同一份文件。最好的，杰夫--由杰夫·达索维奇/NA/安然于2001年5月31日下午5时26分转发--德莱尼·亨特<[dhunter@s-k-w.com>](mailto:dhunter@s-k-w.com)2001年5月30日下午12时29分请回复dhunter：杰夫·达索维奇（电子邮件）<[jdasovic@enron.com>](mailto:jdasovic@enron.com)抄送：主题：Asm Dem计划一个我们手头上的小东西。.. - 装配演示计划. doc

* 1. 根据反馈提示进行QA细化

在这里，我们包括未优化的提示，用于QA细化，以使问题更高质量，以及DSPy MIPROv2 [[55](#_bookmark72)]优化的提示，包括重写的指令和两个引导的fewshot示例。 此提示用于优化步骤，以根据自动生成的反馈提高问题的质量。 这在§B.1中描述的相同端到端优化中进行优化[。](#_bookmark101)

反馈质量保证改进（DSPy优化）（第1部分）

你是一个高风险的团队领导，负责为一家财富500强公司的关键项目提炼问题 您的目标是生成可用于训练AI模型的高质量，客观的问题。 您已收到一封电子邮件、一个与电子邮件相关的问题以及专家对问题质量的反馈。 你的任务是细化问题以解决反馈，而不显著改变其含义，并确保它仍然可以用一句话回答。 您的问题的质量将直接影响项目的成功和公司的声誉。仔细提炼这个问题，并为你所做的改变提供你的理由

-

遵循以下格式。

电子邮件：我们要优化问题以进行反馈的电子邮件：关于问题质量的反馈

推理：让我们一步一步地思考，以便${产生新问题}。我们... 精炼问题：针对反馈的精炼问题

-

电子邮件：主题：FW：授权交易商名单电子邮件：[alterson\_thomas@jpmorgan.com](mailto:alterson_thomas@jpmorgan.com)收件人：[sara. enron.com文件](mailto:sara.shackleton@enron.com)：shackleton-s/all\_documents/11472。

=====================================

你能看一下这个吗

我不想烦你，但我们现在正在进行内部审计，

我们因为这份名单受到了压力

感谢您抽出宝贵的时间。汤姆·艾特森

Thomas P Alterson06/05/200112：18PM

收件人：[Sara. enron.com](mailto:Sara.Shackleton@enron.com)抄送：

主题：授权交易商列表（文档链接未转换）Sara：

你能给我一份安然公司授权交易商的名单吗我们

更新我们的文件，并注意到我们缺少一个ATL的安然。

ATL表格：

(See附加文件：ATL Enron NA Corp.doc）

谢谢你，汤姆·艾特森

- ATL Enron NACorp.doc

=====================================

反馈质量保证改进（DSPy优化）（第2部分）

问：Tom Alterson给出的需要Enron NA Corp授权交易商名单的原因是什么

反馈：问题不客观。 不同的注释者在给定相同的电子邮件作为上下文的情况下以不同的方式回答相同的问题。可能会受益于更清晰。

推理：让我们一步一步地思考，以完善这个问题。 我们需要使问题更加客观和明确。TomAlterson在电子邮件中提到需要EnronNACorp授权交易员名单的原因是“我们正在更新我们的文件，并注意到我们缺少Enron的ATL“然而，这个原因并没有明确说明是主要原因，而是次要原因。 主要原因是在转发的电子邮件中提到的“我们现在正在进行内部审计，我们受到了这份名单的压力。“为了使问题更客观，我们可以重新措辞，把重点放在电子邮件中提到的主要原因。

细化问题：Tom Alterson在内部审计期间需要Enron NA Corp授权交易员名单的主要原因是什么？

-

电子邮件：主题：FW：授权交易商名单电子邮件：[alterson\_thomas@jpmorgan.com](mailto:alterson_thomas@jpmorgan.com)收件人：[sara. enron.com文件](mailto:sara.shackleton@enron.com)：shackleton-s/all\_documents/11472。

=====================================

你能看一下这个吗

我不想烦你，但我们现在正在进行内部审计，

我们因为这份名单受到了压力

感谢您抽出宝贵的时间。汤姆·艾特森

Thomas P Alterson06/05/200112：18PM

收件人：[Sara. enron.com](mailto:Sara.Shackleton@enron.com)抄送：

主题：授权交易商名单（文档链接未转换）Sara：

你能给我一份安然公司授权交易商的名单吗我们

更新我们的文件，并注意到我们缺少一个ATL的安然。

ATL表格：

(See附加文件：ATL Enron NA Corp.doc）

谢谢你，汤姆·艾特森

- ATL Enron NACorp.doc

=====================================

问：TomAlterson在内部审计期间需要EnronNACorp授权交易员名单的主要原因是什么

# 提问提问

这些提示用于回答给定电子邮件上下文的问题，或者在没有基础的情况下产生问题的答案。强迫LLM回答问题而不接地是为了确保问题不太容易猜测或不被流行的LLM记住

# LLM作为一个法官提示

这个提示被用于我们的法学硕士作为判断，以确定是否两个答案是相同的或不同的。法学硕士作为一名法官是以文件为基础的，这有助于它确定特定答案中的额外细节是幻觉还是以事实信息为基础。我们抽样了100个LLM作为法官认为答案匹配的实例，以及100个LLM作为法官认为答案不匹配的实例。作者手动标记这些评估，我们确定法学硕士作为法官

0.98 F1-评分（在两种情况下与人类法官的2个判断不同 这给了我们很高的信心，在使用我们的法学硕士法官thoughout我们的评估。重要的是要注意，我们不是使用LLM作为判断者来进行主观判断，而是确定两个开放式答案是否匹配。 这解释了即使LLM作为判断者可能不可靠时的高准确性[[73]。](#_bookmark90)

# 基于规则的质量评估提示

这些提示用于回答给定电子邮件上下文的问题，或者在没有基础的情况下产生问题的答案。强迫LLM回答问题而不接地是为了确保问题不太容易猜测或不被流行的LLM记住

基于规则的质量评估

给定一封特定的电子邮件，一个关于电子邮件的问题，以及这个问题的答案，判断这个问题的质量。该问题应遵循以下标准：

* + 1. 这个问题应该是具体到这封电子邮件的事实内容换句话说，如果你收到100封电子邮件，你会知道问题是关于这封电子邮件而不是另一封吗？ 诸如“电子邮件”或“信息”之类的短语不够具体，除非有进一步的上下文基础。仅仅是发件人可能还不够，但发件人和有关电子邮件的一些细节可能是足够的。
    2. 问题应该集中在邮件的主要内容上，而不是格式，发件人或收件人。可以使用发件人和收件人作为上下文，但问题不应该是关于他们的。可以询问电子邮件中的人的手机号码或联系方式，因为这是一个现实的问题。
    3. 这个问题应该是客观的，可以用一句话回答 这不应该是一个意见问题，也不需要任何解释。
    4. 这个问题应该是现实的，一个人可能会问他们收到的电子邮件，特别是在专业环境中工作和回忆重要细节的情况
    5. 这个问题不应该要求任何外部知识超出电子邮件本身的内容。
    6. 问题不应该包括计数，数学或任何其他操作。 例如，“发件人提到'重要'这个词多少次？“或“有多少个收件人？”是不允许的。这个问题可以问电子邮件中的一个数字，例如百分比，但它不应该涉及实际计数或计算。

根据这些规则，将问题评为“好”或“坏”。

-

遵循以下格式。

电子邮件：我们要回答的电子邮件问题：我们要回答的关于电子邮件的问题回答：问题的答案

推理：让我们一步一步地思考，以便${产生质量}质量：问题是好（真）还是坏（假）

-

电子邮件：{**email**}问题：{**question**}答案：{**answer**}

推理：让我们一步一步地思考，