RAFT：适应特定领域的语言模型

张天军*
加州大学伯克利分校计算机科学系
美国加利福尼亚州伯克利市，邮编：94720，电子邮箱：{tianjunz}@berkeley.edu

Shishir G. Patil、Naman Jain、Sheng Shen
加州大学伯克利分校计算机科学系
美国加利福尼亚州伯克利市，邮编：94720
{shishirpatil，naman_jain，sheng. s}@berkeley.edu

Matei Zaharia、Ion Stoica、Joseph E. Gonzalez
加州大学伯克利分校计算机科学系
美国加利福尼亚州伯克利市，邮编：94720
{matei,istoica,jegonzal}@berkeley.edu

摘要

在大型文本数据集上预训练大型语言模型（LLMs）已成为标准做法。在将这些模型应用于多种下游应用时，通常会通过基于RAG的提示或微调来向预训练模型中添加新信息。然而，如何最有效地整合这些信息仍然是一个未解之谜。本文介绍了一种名为检索增强微调（RAFT）的新方法，该方法能够提升模型在‘开放书’领域内回答问题的能力。在训练RAFT时，给定一个问题和一组检索到的文档，模型会被训练忽略那些无助于解答问题的文档，即所谓的干扰文档。RAFT通过逐字引用相关文档中的正确序列来帮助回答问题。
这与RAFT的思维链式响应相结合，有助于提高模型的推理能力。在特定领域的RAG中，RAFT持续提升模型在PubMed、HotpotQA和Gorilla数据集上的表现，提供了一种训练后的方法来改进预训练的LLM，使其适应领域内的RAG。

1 介绍

经过大量公共数据的训练，大型语言模型（LLMs）在广泛的通用知识推理任务中取得了显著进展（Brown等人，2020；Wei等人，2022）。然而，越来越多的LLMs被应用于特定领域，支持从特定软件框架的代码补全到特定文档集合（如法律或医学文档）的问题回答等任务。在这些场景中，通用知识推理的重要性相对较低，主要目标是基于给定的文档集提高准确性。实际上，将LLMs适应于特定领域（如。，最近的新闻、企业私有文档或培训截止后构建的程序资源）对于许多新兴应用（Vu等人，2023；Lazaridou等人，2022）至关重要，并且是本研究的重点。

本文研究了以下问题——我们如何在专业领域中将预训练的LLM应用于检索增强生成（RAG）？

在将大语言模型适应特定领域时，我们考虑了以下两种方法：通过检索增强生成（RAG）进行上下文学习和监督微调。基于RAG的方法允许大语言模型在参考文档时

*通讯作者，个人网站：tianjunz.github.io

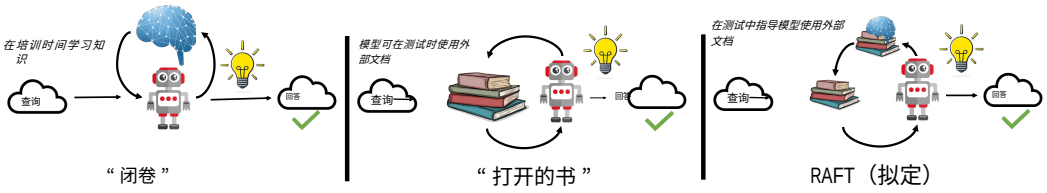

图1：如何最好地准备考试？(a)通过精细调整的方法，考生可以通过直接记忆输入文档或在不参考文档的情况下回答练习题来“学习”。(b)另一种方法是上下文检索，这种方法未能充分利用固定领域的学习机会，相当于没有准备就参加开卷考试。相比之下，我们的方法(c) RAFT通过在模拟的不完美检索环境中参考文档，利用问题-答案对进行精细调整——从而有效地为开卷考试做准备。

回答问题。然而，基于RAG的上下文学习方法未能利用固定领域设置提供的学习机会和早期访问测试文档的机会。或者，监督微调提供了学习文档中更普遍模式的机会，从而更好地与最终任务和用户偏好对齐（Zhou等，2023）。然而，现有的基于微调的方法要么在测试时未能利用文档（不整合RAG)，要么在训练过程中未能考虑到检索过程中的不完美。

我们可以将现有的上下文检索方法比作未做准备的开卷考试。而现有的微调方法则是通过直接‘记忆’Xiong等人（2023）的输入文档，或在不参考文档的情况下回答Wang等人（2022）的练习题来‘学习’。尽管这些方法利用了领域内学习的优势，但它们未能为考试的开放性做好准备。

本文研究了如何将指令微调（IFT）与检索增强生成（RAG）相结合。我们提出了一种新的适应策略——检索增强微调（RAFT）。RAFT特别针对了微调大型语言模型（LLM）的挑战，旨在在提升领域内检索与回答（RAG）性能的同时，融入领域知识。RAFT的目标不仅在于通过微调使模型能够学习特定领域的知识，还在于确保模型对干扰信息的鲁棒性。这是通过训练模型理解问题（提示）、检索到的特定领域文档和正确答案之间的关系来实现的。回到开放书籍考试的比喻，我们的方法类似于通过识别相关和不相关的检索文档来备考。

在RAFT中，我们训练模型从文档(s)（D*）中回答问题(Q)，生成答案（A*），其中A*包括思路推理Wei等人（2022）；人类假设（2023），并在存在干扰文档（Dk）的情况下进行。我们在第3节解释了方法论，并在第5节分析了训练和测试时对干扰文档数量(k)的敏感性。无论是否使用RAG，RAFT在PubMed Dernoncourt & Lee（2017）、HotPot QA Yang等人研究中均优于监督微调。（2018），以及HuggingFace Hub、Torch Hub和Tensorflow Hub Gorilla数据集Patil等人（2023），提出了一种新颖而简单的技术，以改进预训练的LLM用于领域内RAG。我们的代码可在https://github.com/ShishirPatil/gorilla获取。

2 开放考试的LLM

为了更好地理解我们的目标，我们扩展了在准备考试的现实世界环境中训练LLM的类比。

闭卷考试通常指的是LLM的考试
在此期间，无法访问任何其他文件或参考资料来回答这些问题

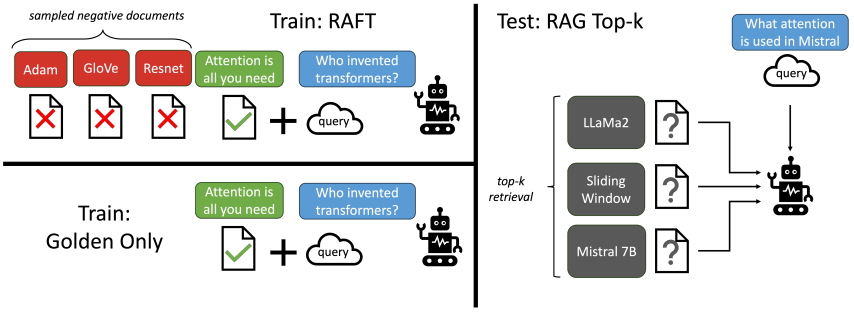

图2：我们的RAFT方法概览。左上角的图展示了我们如何将大语言模型适应于从一组正文档和干扰文档中阅读解决方案，这与标准RAG设置形成对比，在标准RAG设置中，模型是基于检索器输出进行训练的，这是记忆和阅读的混合体。在测试时，所有方法都遵循标准RAG设置，提供上下文中的前k个检索到的文档。

考试。对于LLM，这相当于将LLM用作聊天机器人的场景。在这种场景中，LLM利用预训练和监督微调过程中积累的知识来响应用户的提示。

开放书籍考试。相比之下，开放书籍考试的设置类似于法律硕士（LLM）可以参考外部信息源（如网站或书籍章节）的情况。在这种情况下，LLM通常与检索器配合使用，检索器会获取‘k’篇文档（或文档的特定部分），并将这些内容附加到用户的查询中。只有通过这些检索到的文档，LLM才能访问“领域特定信息”。因此，我们认为，在这些设置中，LLM的性能很大程度上取决于检索器的质量以及检索器识别最相关信息的准确性。

领域特定开放考试在本文中，我们关注的是比一般开放考试更狭窄但越来越受欢迎的领域，即领域特定开放考试。在这里，我们事先知道LLM将要测试的领域。LLM可以使用来自该特定领域的所有信息来回应用户的提示，这些信息是经过微调的。领域特定的例子包括企业文档、属于某个组织的代码库等。在所有这些场景中，LLM将用于回答问题，而这些问题的答案可以在一系列文档中找到。检索技术本身对机制的影响很小（尽管可能会影响准确性）。本文研究了特定领域的开放书设置，以及如何将预训练的LLM适应到这个特定领域，包括如何使其对检索到的文档和干扰物数量的变化更具鲁棒性。

3 木筏

在本节中，我们介绍了一种新的训练方法——RAFT，用于特定领域的开放书考试。首先，我们介绍了监督微调这一经典技术，接着分享了实验的关键发现。随后，我们介绍了RAFT，这是一种改进的通用指令调优方法。最后，我们概述了后续章节中将要进行的实验。

监督微调

考虑监督微调（SFT）设置下的问答数据集。该设置包括一个数据集(D)，从中可以推导出或已知一组问题(Q)及其对应答案(A)。在传统的SFT设置中，模型根据其知识——这些知识要么是在预训练阶段获得的，要么是在SFT训练阶段获得的——来提高回答问题的能力。经过这样训练的模型还可以

在测试时使用检索增强生成（RAG）设置，其中可以在提示中引入额外的文档以帮助模型回答问题。这可以表示如下：

{训练：Q→A}，{0-shot推理：Q→A}，{RAG推理：Q + D→A}

RAFT：检索增强微调（RAFT）提出了一种新颖的方法来准备微调数据，以适应特定领域的开放问答设置，相当于领域内的RAG。在RAFT中，我们准备训练数据，使得每个数据点包含一个问题(Q)、一组文档（Dk），以及一个来自其中一个文档（D *）的链式答案（A*）。我们区分了两种类型的文档：‘ 黄金 ’文档（D *），即可以从其中推导出问题答案的文档；以及‘ 干扰 ’文档（Di），这些文档不包含与答案相关的信息。作为实现细节，‘ 黄金 ’文档不必是单个文档，可以是多个文档，如Yang等人（2018）在HotpotQA中的情况。然后，对于数据集中问题（qi）的P比例部分，我们保留黄金文档（d*i）以及干扰文档（dk?1）。对于数据集中问题（qi）的（1?P）比例部分，我们不包含黄金文档，仅包含干扰文档（dk）。接着，我们使用标准监督训练（SFT）技术微调语言模型，使其能够从提供的文档和问题中生成答案。图2展示了RAFT的高层次设计原则。

我们证明了我们的RAG方法能够训练模型在所训练的文档集上表现更佳，即在领域内。通过在某些情况下移除黄金文档，我们促使模型记忆答案而不是从上下文中推导答案。RAFT的训练数据如下所示，图3中可以看到一个示例训练数据：

数据的P%部分：Q + D* + D?+D?+···+D?→A*

（1?P）%的数据：Q+D1+D2+···+dk→A*

随后，对于测试场景，模型将Q和RAG管道检索到的前k个文档提供给模型。请注意，RAFT与使用的检索器是独立的。

提高训练质量的关键因素之一是生成推理过程，例如思维链，以解释提供的答案。RAFT方法与此类似：我们证明了创建完整的推理链，并且明确引用来源可以增强模型回答问题的准确性。图3展示了这一设置。以这种方式生成训练数据，包括向模型呈现问题、上下文和经过验证的答案，然后要求其形成一个适当引用原始上下文的推理链。

在我们的实验中，对于所有数据集，我们使用上述技术生成答案。需要注意的是，Gorilla APIBench数据集的每个答案都包含了推理过程。图3展示了生成步骤的一个例子，其中详细的推理回答不仅包括了引用原始上下文的引号（##begin_quote##和##end_quote##），还详细解释了如何基于这些引用得出结论。我们在实验部分展示了添加详细推理段落可以显著提升模型性能的效果。

4 评价

我们设计实验，旨在评估RAFT与多种基线模型相比的表现。研究发现，经过微调的LlaMA-2版本——RAFT-7B，在阅读和提取领域内文档信息方面表现更佳，优于特定领域的微调模型和结合RAG的通用模型。此外，我们还通过消融实验展示了链式思维响应对模型学习的重要性。在本节中，我们将首先介绍实验中使用的所有数据集，然后列出所有用于对比的基线模型及微调技术。

问题：Oberoi家族是某酒店公司的成员，该公司总部位于哪个城市？

背景：[奥贝罗伊家族是一个印度家族，因涉足酒店业而闻名，特别是通过奥贝罗伊集团]...[它位于雅加达市中心，靠近梅加库宁安，毗邻姐妹酒店JW万豪酒店。该酒店由丽思卡尔顿酒店公司运营。综合体包含两座塔楼，分别是一座酒店和艾尔朗加公寓]...[奥贝罗伊集团是一家酒店公司，总部设在德里。]

说明：根据上面的问题、背景和答案，提供一个合乎逻辑的
回答的理由。请使用以下格式：##理由：{reason}##答案： {answer}。

CoT回答：##原因：文件##begin_quote##奥贝罗伊家族是一个印度家族，因参与酒店业而闻名，特别是通过奥贝罗伊集团。##end_quote##这表明奥贝罗伊家族参与了奥贝罗伊集团，文件##begin_quote##奥贝罗伊集团是一家总部位于德里的酒店公司。##end_quote##则确定了奥贝罗伊集团的总部位置。因此，Oberoi家族是总部位于德里的酒店公司的一部分。##答案：德里

图3：RAFT提示帮助LLM评估其自动生成的推理和答案，并将其与正确的推理和答案进行对比。LLM被提示识别其推理中的错误并提取关键见解以改进。该图具体表示RAFT算法中的“ 生成解释 ”步骤（第3节）。

表1：RAFT提高了所有专业领域的RAG性能：在PubMed、HotPot、HuggingFace、Torch Hub和Tensorflow Hub上，我们发现领域特定微调显著提升了基础模型的性能，无论是否使用RAG，RAFT始终优于现有的领域特定微调方法。这表明需要在上下文中训练模型。我们将模型与LLaMA微调方案进行比较，并提供GPT-3.5作为参考。

	PubMed	火锅	拥抱脸	火炬中心	TensorFlow
GPT-3.5 + RAG	71.60	**41.5**	29.08	60.21	65.59
LLaMA2-7B	56.5	0.54	0.22	0	0
LLaMA2-7B + RAG	58.8	0.03	26.43	08.60	43.06
差示分光荧光计	59.7	6.38	61.06	84.94	86.56
DSF +拉格	71.6	4.41	42.59	82.80	60.29
RAFT（LLaMA2-7B）	**73.30**	35.28	**74.00**	**84.95**	**86.86**

数据集在我们的实验中，我们使用以下数据集来评估我们的模型和所有基线。我们选择了这些数据集来代表流行且多样的领域，包括维基百科、编码/API文档以及医学文档中的问答。自然问题（NQ）Kwiatkowski等人（2019年）、Trivia QA Joshi等人（2017年）和HotpotQA Yang等人（2018年）是基于维基百科的开放领域问答，主要关注常识（如电影、体育等）。HuggingFace、Torch Hub和TensorFlow Hub来自Patil等人（2023年）在Gorilla论文中提出的APIBench。这些基准测试衡量如何根据文档生成正确、功能性和可执行的API调用。PubMed QA Jin等人（2019年）是一个专门针对生物医学研究问答的数据集，主要集中在基于给定文档集回答医学和生物学问题。我们将

需要强调的是，（NQ、Trivia QA和HotpotQA）是相对通用的领域，而后两个领域是基于特定领域的文档。

基线我们在实验中考虑了以下基线：

- ? LlaMA2-7B-chat模型，使用0-shot提示：这是QA任务中常用的指令微调模型，我们提供清晰的书面说明，但没有参考文档。
- ? LlaMA2-7B-chat模型与RAG（Llama2 + RAG）：与之前的设置类似，但这里我们加入了参考文档。这是处理领域特定问答任务时常用的一种技术。
- ? 领域特定微调（DSF）：采用无样本提示的标准化监督微调方法，不包含上下文文档。我们发现，这种方法主要有助于调整模型的回答风格，并熟悉领域背景。
- ? 使用RAG进行领域特定微调（DSF + RAG）：利用RAG为领域特定的微调模型配备外部知识。因此，对于模型未知的“知识”，它仍然可以参考上下文。

4.1 结果

利用上述数据集和基线模型，我们评估了RAFT模型，并在表1中展示了RAFT的有效性。结果显示，RAFT在性能上显著优于基线模型。与基础的Llama-2指令调优模型相比，结合RAG的RAFT在信息提取和对干扰项的鲁棒性方面表现更为出色。在Hotpot QA任务上，其性能提升了35.25%；在Torch Hub评估中，提升了76.35%。与DSF在特定数据集上的表现相比，我们的模型在利用上下文解决问题方面表现更佳。RAFT在处理Hotpot和HuggingFace数据集的任务上表现尤为突出（Hotpot任务上提升了30.87%，HuggingFace任务上提升了31.41%）。值得注意的是，对于PubMed QA，由于这是一个简单的“是/否”问题，与DSF + RAG相比，我们的模型并未显示出显著的性能提升。即使与更大、更强大的GPT-3.5模型相比，RAFT依然展现出显著的优势。

总体而言，无论是否使用RAG，LLaMA-7B模型的表现都较差，因为其回答方式与真实情况不符。通过应用领域特定的微调，我们显著提升了其性能。这一过程使模型能够学习并采用合适的回答风格。然而，将RAG引入特定领域微调（DSF）模型并不一定会带来更好的结果。这可能表明模型在上下文处理和从中提取有用信息方面缺乏训练。通过结合我们的方法RAFT，我们不仅训练模型使其回答风格与所需匹配，还提高了其文档处理能力。因此，我们的方法优于所有其他方法。

4.2 CoT的影响

我们还进行了分析，以评估思路链方法在提升模型性能方面的有效性。如表2所示，仅仅提供问题的答案可能并不总是足够的。这种方法可能导致损失迅速下降，从而使模型开始过拟合。结合推理链不仅能够引导模型找到答案，还能丰富模型的理解，从而提高整体准确性并防止过度拟合到简洁的答案。在我们的实验中，整合思路链显著增强了训练的鲁棒性。我们使用GPT-4-1106生成我们的思路链提示，并在图3中包含一个我们使用的提示示例。

4.3 定性分析

为了展示RAFT相较于领域特定微调（DSF）方法的潜在优势，我们在图4中提供了一个对比示例。该示例定性地

表2：基于思路链的消融实验：包含和不包含CoT的RAFT数量。不同数据集上的结果表明，加入CoT可以显著提升微调模型的性能，在Hotpot QA和HuggingFace数据集上分别提高了9.66%和14.93%。

	PubMed	火锅QA	HuggingFace	火炬中心	TensorFlow
RAFT与CoT合作	68.30	25.62	59.07	**86.56**	83.21
木筏	**73.30**	**35.28**	**74.00**	84.95	**86.86**

展示了DSF模型在面对一个询问编剧身份的问题时变得困惑的情景。它没有提供正确的姓名，而是错误地引用了该编剧所写的其中一部电影。相比之下，RAFT模型准确地回答了问题。这种差异表明，仅用问答对训练模型可能会削弱其从提供的文档中提取相关上下文的能力。这一对比强调了在训练数据集中结合标准指令调优和上下文理解的重要性，以保持并增强模型有效处理文本的能力。

HotPot质量保证部

问题：哪位编剧曾为尼古拉斯·凯奇和蒂娅·莱奥尼主演的电影《进化》[0/1879]写过剧本？
文件：... David 魏斯曼是一名编剧和导演。他的电影作品包括《家庭男人》（2000年）、《进化》（2001年），和《当我在罗马》（2010年）。
...
《家庭男人》是一部2000年的美国浪漫喜剧剧情片，由布雷特·拉特纳执导，大卫·戴蒙德和大卫·韦斯曼编剧，尼古拉斯主演。笼子和茶莱奥尼...

RAFT：##原因：电影《进化》的编剧是大卫·韦斯曼，该片由尼古拉斯·凯奇和蒂娅·莱奥尼主演。这一信息在参考文献中提到，大卫·韦斯曼作为编剧，曾参与《家庭男人》（2000年）、《进化》（2001年）和《罗马假日》（2010年）等影片的创作。因此，《进化》的编剧是大卫·韦斯曼。##答案：大卫 魏斯曼√
DSF：是 家庭 男X

图4：RAFT与DSF的对比：在HotPot QA数据集上，我们可以看到当问题问的是谁是编剧时，DSF模型从上下文中提取了错误的信息，并回答了一个电影名称。而RAFT则成功地获得了准确的结果。

4.4 我们是否应该始终使用黄金上下文来训练用于RAG的LLM？

在探讨大型语言模型（LLM）是否应始终使用黄金上下文进行检索增强生成（RAG）训练时，我们关注一个关键问题：训练数据中应包含多少比例（p%）的黄金文档？直观上，人们可能会认为为了有效训练阅读和从上下文中提取信息（例如，RAG任务），训练过程中应始终包含黄金文档（P = 100%）。然而，我们的研究结果挑战了这一假设：在上下文中不包含黄金文档的情况下，仅使用部分训练数据（P = 80%）似乎也能提升模型在RAG任务上的表现。

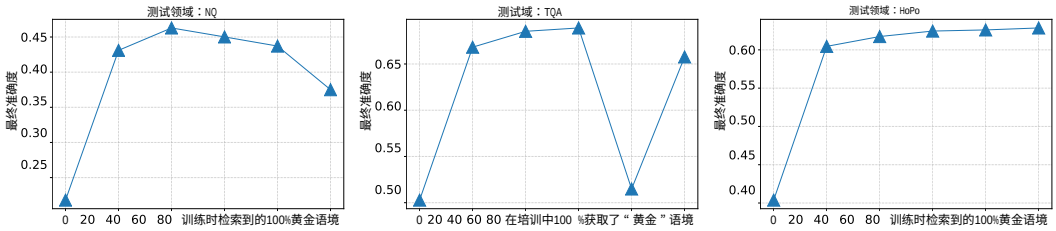

图5：涉及多少个黄金文档？我们研究了超参数P%，它表示训练数据中有多少比例包含黄金文档。NQ、TQA和HotpotQA的结果表明，混合一些黄金文档未出现在上下文中的数据有助于领域内RAG。

图5展示了我们对超参数P%的研究，该参数表示训练实例中应包含黄金文档的比例。我们发现最优比例因数据集而异，P%范围从40%、60%到100%不等。这表明，在没有正确上下文的情况下训练大语言模型有时对下游任务，即回答与文档相关的问题是有益的。在我们的训练设置中，除了黄金文档外，还加入了四个干扰文档。在测试时，我们保持这一格式，即黄金文档旁边附有四个干扰文档。研究发现，在特定领域的RAG任务中，包含一定比例的没有黄金文档的训练数据是有益的。

5 RAFT推广至Top-K RAG

我们现在探讨另一个重要问题：在评估时加入top-k RAG结果，RAFT中干扰文档的数量如何影响模型的性能？先前的研究指出，大型语言模型（LLM）对无关文本的敏感性（参见Shi等，2023a；Weston和Sukhbaatar，2023；Liu等，2023）。这一问题对于+ RAG的LLM尤为关键，因为top-k RAG常常在测试阶段使用，以确保高召回率。在这种情况下，模型需要具备识别并忽略无关内容的能力，专注于相关的信息。

5.1 使模型对前K个RAG具有鲁棒性

为了应对增强大型语言模型（LLM）在检索管道中筛选无关文本的能力的挑战，我们的分析表明，仅使用黄金文档（高度相关）进行训练可能会无意中削弱模型识别和忽略无关信息的能力。为此，我们的算法RAFT采用了一种策略，将黄金文档与一些无关文档混合使用。这一方法促使我们研究在整个训练过程中应包含的理想比例的干扰（无关）文档，并评估这种训练方法在测试阶段适应不同文档量的能力。我们的目标是优化相关和无关信息之间的平衡，以提高模型识别和利用相关内容的效率。请注意，第4.4节探讨了训练数据中应包含多少百分比的干扰项，而本节则研究测试时的情景。

使用干扰文档进行训练为了增强大语言模型对检索文档中无关文本的鲁棒性，我们采用了一种微调方法，该方法结合了黄金（高度相关）文档和干扰（无关）文档。模型在不同数量的干扰文档下进行了训练，但始终使用来自检索器的前3个文档进行评估——不要与p混淆。我们的研究结果如图6所示，仅使用黄金文档进行微调时，性能通常不如包含更多干扰文档的配置。从图中可以看出，自然问题的表现更好。

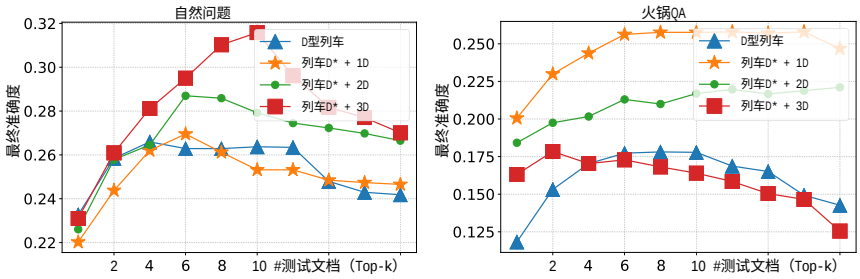

图6：测试时间文档变化：为了分析RAFT对不同数量测试时间文档的鲁棒性，我们研究了三个领域——NQ、Trivia QA和HotPot QA。在NQ中，我们发现使用4个文档训练可以达到最佳性能，而Trivia QA和HotPot QA则分别变为3个和2个。然而，我们发现仅使用黄金文档训练会导致性能较差。

使用D * + 3D进行训练，而D * + 1D则是带有Hotpot QA的文档。这一见解对我们算法RAFT特别有益。在我们的实验中，我们始终采用一种训练设置，包括一个黄金文档和四个干扰文档。

推广到可变数量的测试时间文档。我们扩展了研究，以考察不同数量的测试时间文档对模型性能的影响。具体来说，我们的实验集中在评估不同数量干扰文档训练的模型如何应对测试时文档数量的变化。图6所示的结果证实，在训练过程中加入干扰文档确实使模型更能适应测试中遇到的文档数量波动。这种即使在测试文档数量变化的情况下仍能保持一致性能的能力进一步验证了我们方法RAFT的稳健性。这一发现强调了精心校准的训练环境对于准备模型应对现实世界中可能遇到的各种场景的重要性。

6 相关作品

检索增强语言模型（RALMs）通过集成一个检索模块，该模块从外部知识库中获取相关信息，显著提升了各种自然语言处理任务的性能，包括语言建模(Guu等人，2020；Borgeaud等人，2022；Khandelwal等人，2019；Shi等人，2023d；Lin等人，2023b；Shi等人，2023c；Asai等人，2023；Xu等人，2023；Wang等人。，2023)和开放领域问答（Izacard等，2023；Lewis等，2020）。例如，Atlas（Izacard等，2023）使用检索器微调T5模型，将文档视为潜在变量，而RETRO（Borgeaud等，2022）则修改仅解码器架构，加入检索到的文本，并从头开始预训练。kNN-LM（Khandelwal等，2019）在推理时插值于语言模型的下一个标记分布和从检索到的标记计算出的分布之间。（Shi等，2023d；Ram等，2023）假设可以黑盒访问语言模型，将其与现成或微调的检索器结合使用。

记忆大型神经语言模型的一个关键问题是它们是否真正“理解”文本（Feldman，2020；Power等，2022），还是仅仅依赖于表面模式的记忆（Carlini等，2019；Tanzer等，2022）。（Feldman，2020；Carlini等，2019；2022）开发了方法来量化神经模型中记忆的程度。（Brown等，2020；Power等，2022；Liu等，2022)进一步探讨了记忆如何影响模型的泛化能力。(Carlini等，2021；Shi等。2023b的研究表明，语言模型能够记忆并复述训练数据，这引发了重大的隐私问题（Kandpal等，2022；Pan等，2020）。

最近，多篇论文探讨了通过微调预训练的大型语言模型（LLM）以提升其在阅读理解与问答（RAG）任务中的表现（林等，2023a；王等，2023；徐

（等，2023；刘等，2024）。这些研究主要集中在构建用于RAG的微调数据集，并训练模型以在这些任务上表现优异。特别是，在他们的实验设置中，测试时使用的领域或文档可能与训练时不同；而我们的研究则关注一个相反的情况，即只在相同的文档集上测试大型语言模型。

7 结论

RAFT是一种训练策略，旨在提升模型在特定领域内回答问题的能力，在“开放问答”环境中表现更佳。我们强调了几个关键的设计决策，例如与干扰文档一起训练模型、组织数据集以确保部分文档在其上下文中缺乏黄金文档，以及采用思路链方式直接引用相关文本来构建答案。我们在PubMed、HotpotQA和Gorilla API基准上的评估突显了RAFT的巨大潜力。

参考文献

人为因素。为克劳德的长期上下文窗口进行提示工程。2023年。

A.Asai、Z.Wu、Y.Wang、A.Sil和Hajishirzi，H.自我反思：通过自我反思学习检索、生成和批评。arXiv预印本arXiv：2310.11511,2023。

Borgeaud，S.，Mensch，A.，Hoffmann，J.，Cai，T.，Rutherford，E.，Millican，K.，Van Den Driessche，G. B.，Lespiau，J.-B.，Damoc，B.，Clark，A.,等.通过检索数万亿个标记来改进语言模型.国际机器学习会议,第2206-2240页.PMLR,2022.

Brown，T.，Mann，B.，Ryder，N.，Subbiah，M.，Kaplan，J. D.，Dhariwal，P.，Neelakantan，A.，Shyam，P.，Sastry，G.，Askell，A.,等.语言模型是少样本学习者.神经信息处理系统进展,33:1877-1901,2020.

Carlini，N.，Liu，C.，Erlingsson，U.，Kos，J.和Song，D.《秘密分享者：评估和测试神经网络中的非预期记忆》。第28届USENIX安全研讨会（USENIX Security 19），第267-284页，2019年。

Carlini，N.，Tramer，F.，Wallace，E.，Jagielski，M.，Herbert-Voss，A.，Lee，K.，Roberts，A.，Brown，T.，Song，D.，Erlingsson，U.,et al.从大型语言模型中提取训练数据.收录于第30届USENIX安全研讨会(USENIX Security 21),第2633-2650页,2021年.

卡林尼，N.，伊波利托，D.，雅吉尔斯基，M.，李，K.，特拉默，F.，张，C.，《跨神经语言模型的记忆量化》。收录于《第十一届国际学习表示会议论文集》，2022年。

德诺库尔，F.和李，J. Y. Pubmed 200k rct：一个用于医学摘要中顺序句子分类的数据集。arXiv预印本arXiv：1710.06071,2017。

菲尔德曼，V。学习需要记忆吗？一个关于长尾的短故事。在第52届ACM SIGACT计算理论年会论文集，第954-959页，2020年。

古，K.，李，K.，董，Z.，帕苏帕特，P.，和张，M.检索增强语言模型预训练。国际机器学习会议，第3929-3938页。PMLR，2020。

伊扎卡尔德，G.，刘易斯，P.，洛梅利，M.，霍塞尼，L.，佩特罗尼，F.，希克，T.，迪维迪-尤，J.，乔林，A.，里德尔，S.，格拉夫，E.，《Atlas：基于检索增强语言模型的少样本学习》。《机器学习研究杂志》，24（251）：1－43,2023年。网址http://jmlr.org/papers/v24/23-0037.html。

金，Q.，丁格拉，B.，刘，Z.，科恩，W. W.，和陆，X. Pubmedqa：一个用于生物医学研究问题解答的数据集。arXiv预印本arXiv：1909.06146,2019。

乔希，M.，崔，E.，韦尔德，D. S.，和泽特勒莫耶，L. Triviaqa：一个用于阅读理解的大规模远程监督挑战数据集。arXiv预印本arXiv：1705.03551,2017。

Kandpal，N.、Wallace，E.和Raffel，C.去重训练数据可降低语言模型中的隐私风险。在国际机器学习会议上，第10697-10707页。PMLR，2022。

Khandelwal，U.、Levy，O.、Jurafsky，D.、Zettlemoyer，L.和Lewis，M.，总括通过记忆实现：最近邻语言模型。arXiv预印本 *arXiv:1911.00172*，2019.

Kwiatkowski，T.，Palomaki，J.，Redfield，O.，Collins，M.，Parikh，A.，Alberti，C.，Epstein，D.，Polosukhin，I.，Devlin，J.，Lee，K.，等.自然问题:问答研究的基准.计算语言学协会会刊,7:453-466,2019.

拉扎里杜，A.，格里博夫斯卡娅，E.，斯托科维茨，W.，和格里戈列夫，N.通过少量提示增强互联网语言模型以实现开放领域问答。arXiv预印本arXiv：2203.05115,2022。

Lewis，P.、Perez，E.、Piktus，A.、Petroni，F.、Karpukhin，V.、Goyal，N.、Kuttler，H.、Lewis，M.、Yih，W.-t.、Rocktaschel，T.等。知识密集型nlp任务的检索增强生成。神经信息处理系统进展，33：9459-9474,2020。

林，X. V.，陈，X.，陈，M.，石，W.，洛梅利，M.，詹姆斯，R.，罗德里格斯，P.，卡恩，J.，西尔瓦西，G.，刘易斯，M.等. Ra-dit：检索增强的双指令调优. arXiv预印本 *arXiv:2310.01352*，2023a.

林X. V.、陈X.、陈M.、石W.、洛梅利M.、詹姆斯R.、罗德里格斯P.、卡恩J.、西尔瓦西G.、刘易斯M.等. Ra-dit：检索增强型双指令调优. arXiv预印本 *arXiv:2310.01352*，2023b.

刘，N. F.，林，K.，休伊特，J.，帕拉贾佩，A.，贝维拉夸，M.，佩特罗尼，F.，和梁，P.失落于中间：语言模型如何使用长上下文。arXiv预印本arXiv：2307.03172,2023。

刘，Z.，基图尼，O.，诺尔特，N. S.，米肖，E.，特格马克，M.和威廉姆斯，M.《理解悟解：一种有效的表征学习理论》。《进展》神经信息处理系统，35：34651-34663,2022。

刘，Z.，平，W.，罗伊，R.，徐，P.，肖伊比，M.，和卡坦扎罗，B. Chatqa：构建gpt-4级别的对话问答模型。arXiv预印本arXiv：2401.10225,2024。

潘X.，张M.，季S.，杨M.通用语言模型的隐私风险。2020年IEEE安全与隐私研讨会（SP），第1314-1331页。IEEE，2020。

帕蒂尔，S. G.，张，T.，王，X.，和冈萨雷斯，J. E.《大猩猩：与大规模API相连的大语言模型》。arXiv预印本arXiv：2305.15334,2023。

Power，A.，Burda，Y.，Edwards，H.，Babuschkin，I.，和Misra，V.《Grokking：在小型算法数据集上超越过拟合的泛化能力》。arXiv预印本arXiv：2201.02177,2022年。

Ram，O.、Levine，Y.、Dalmedigos，I.、Muhlgay，D.、Shashua，A.、Leyton-Brown，K.和Shoham，Y.在上下文中检索增强的语言模型。arXiv预印本arXiv：2302.00083,2023。

石福，陈曦，米斯拉，斯凯尔斯，多汉，奇，夏尔利，周。大型语言模型容易被无关的上下文分散注意力。国际机器学习会议，第31210-31227页。PMLR，2023a。

石伟，阿吉斯，夏明，黄勇，刘东，布雷文斯，陈德，泽特洛莫耶，检测大型语言模型的预训练数据。arXiv预印本arXiv：2310.16789,2023b。

施伟、闵思、洛梅利、周晨、李敏、林薇、史密斯、泽特莫耶、叶思和刘易斯。情境预训练：超越文档边界的语言建模。
arXiv预印本arXiv：2310.10638,2023c。

石伟，闵思，安永，西尾，詹姆斯，刘易斯，泽特洛莫耶，以及叶伟涛。Replug：检索增强的黑盒语言模型。arXiv预印本arXiv：2301.12652,2023年。

Tanzer，M.，Ruder，S.和Rei，M.预训练语言模型中的记忆与泛化。在第60届计算语言学协会年会论文集（第1卷：长篇论文），第7564-7578页，2022年。

Vu，T.，Iyyer，M.，Wang，X.，Constant，N.，Wei，J.，Wei，J.，Tar，C.，Sung，Y.-H.，Zhou，D.，Le，Q.等. Freshllms：使用搜索引擎增强刷新大型语言模型。
arXiv预印本arXiv：2310.03214,2023。

王斌，平伟，麦卡菲，刘鹏，李斌，肖伊比，卡坦扎罗。Instructretro：检索后指令调优——增强预训练。arXiv预印本arXiv：2310.07713,2023。

王一，科迪，米什拉，刘，史密斯，哈沙比，和哈吉希尔齐，自我指导：将语言模型与自动生成的指令对齐。arXiv预印本
arXiv:2212.10560, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D.,et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems,35:24824-24837,2022.

韦斯顿，J.和苏赫巴塔尔，S.《系统2注意力（你可能也需要）》。arXiv预印本arXiv：2311.11829,2023年。

Xiong, W., Liu, J., Molybog, I., Zhang, H., Bhargava, P., Hou, R., Martin, L., Rungta, R., Sankararaman, K. A., Oguz, B.,et al.基础模型的有效长上下文扩展. arXiv预印本arXiv:2309.16039,2023.

徐鹏，平伟，吴翔，麦卡菲，朱晨，刘振，苏布拉马尼安，巴赫图里娜，肖伊比，卡坦扎罗。检索与长上下文大语言模型的结合。arXiv预印本arXiv：2310.03025,2023。

杨，Z.，齐，P.，张，S.，本吉奥，Y.，科恩，W. W.，萨拉赫丁诺夫，R.，和曼宁，C. D. Hotpotqa：一个用于多样化、可解释的多跳问答的数据集。arXiv预印本
arXiv:1809.09600, 2018.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L.,等.利马:对齐时少即是多.arXiv预印本arXiv:2305.11206,2023.