动态和参数化的检索增强生成

Weihang Su
swh22@mails.tsinghua.edu.cn
DCST, Tsinghua University
Beijing 100084, China

Qingyao Ai
aiqy@tsinghua.edu.cn
DCST, Tsinghua University
Beijing 100084, China

Jingtao Zhan
jingtaozhan@gmail.com
DCST, Tsinghua University
Beijing 100084, China

Qian Dong
dq22@mails.tsinghua.edu.cn
DCST, Tsinghua University
Beijing 100084, China

Yiqun Liu
yiqunliu@tsinghua.edu.cn
DCST, Tsinghua University
Beijing 100084, China

Abstract

检索增强生成（RAG）已成为为大型语言模型（LLM）配备外部知识的基础范式，在信息检索和知识密集型应用中起着关键作用。然而，传统的 RAG 系统通常采用静态检索后生成的管道，并依赖于上下文中的知识注入，这对于需要多跳推理、自适应信息访问和更深层次整合外部知识的复杂任务可能不是最优的。受到这些限制的启发，研究界已经超越了静态检索和上下文中的知识注入。在新兴的研究方向中，本教程深入探讨了 RAG 的两个快速增长且互补的研究领域：动态 RAG 和参数化 RAG。动态 RAG 在 LLM 的生成过程中自适应地决定何时以及检索什么，从而实现对 LLM 不断变化的信息需求的实时适应。参数化 RAG 重新思考了检索到的知识应该如何注入到 LLM 中，从输入层次的知识注入转变为参数层次的知识注入，以提高效率和效果。本教程提供了这些新兴研究领域中最新进展的全面概述，并分享了理论基础和实践见解，以支持和激发 RAG 的进一步研究。

ACM Reference Format:
Weihang Su, Qingyao Ai, Jingtao Zhan, Qian Dong, and Yiqun Liu. 2025. 动态和参数化的检索增强生成. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3726302.3731692

1

封面信息 标题: 动态和参数化检索增强生成

长度：半天（3 小时）

格式：这是一个半天（3 小时）的讲座式教程，包括计划的休息时间。将会现场进行，至少有两位讲师计划亲自参加会议。

目标受众：中级。本教程适合于有信息检索或自然语言处理经验，并对 IR 技术与大型语言模型的交叉感兴趣的研究人员、从业者和学生。特别适合那些希望了解检索增强生成（RAG）和大型语言模型（LLMs）中知识注入最新进展的人士。

先决知识：参加者需要对基于 Transformer 的大型语言模型 [? ? ? ? ?] 有基本的了解，包括注意力机制等概念。此外，熟悉稀疏 [? ?] 和密集检索 [? ? ? ?] 的基础知识将会有所帮助。

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3731692

之前的演讲：本教程之前未曾展示过。

陈述者：

Weihang Su [1] [?] 是清华大学计算机科学与技术系的博士生，由刘奕群教授指导。他的研究方向是面向检索的生成（RAG）和大语言模型的知识注入。他在包括 AAAI、ACL、EMNLP、WebConf、SIGIR 和 TOIS 在内的顶级会议和期刊上发表了论文。他担任 SIGIR、WebConf、EMNLP、NeurIPS、ICLR、ICML、ACL 和 CIKM 等主要会议的审稿人或程序委员会成员。

清要 · 艾 [2] [?] 是清华大学计算机科学与技术系的副教授。他的研究位于信息检索和生成型人工智能的交叉点，重点是检索/排序优化和检索增强生成。他在马萨诸塞大学阿默斯特分校获得博士学位，导师是 W. Bruce Croft 教授。他曾担任多个学术职务，包括 SIGIR-AP 2023 的大会共同主席、NTCIR-18 的程序共同主席，目前是 ACM TOIS 的副编辑。

展景涛 [3] [?] 是清华大学的一名博士生，由马少平教授和刘奕群教授指导。他的研究重点是高效的信息检索和利用用户交互日志改进 AIGC 系统。他在顶级会议上发表过论文，并获得多项荣誉，包括 WSDM 2022 和 SIGIR 2024 的最佳论文奖。他还担任 SIGIR、ACL 和 The Web Conference 等领先会议的审稿人。

2 动机与概述

大型语言模型（LLMs）在自然语言处理（NLP）和信息检索（IR）任务方面展示了显著的进展，尽管如此，它们面临着诸如幻觉、知识过时和领域适应等挑战。为了应对这些挑战，检索增强生成（RAG）作为一种有前途的解决方案应运而生，通过提供外部知识的访问权来补充 LLMs。传统的 RAG 范式采用一种简单的先检索后生成的方法：一个外部的检索器或一个更复杂的检索系统根据用户的查询检索相关文档，然后这些文档被附加到 LLM 的上下文输入中，随后生成的响应以检索到的证据为基础。这个范式已经在各种知识密集型应用中被证明是有效的，并在实践中被广泛采用。

2.1 标准 RAG 的局限性

在标准的 RAG 中，检索是在生成过程之前进行的，基于最初的用户查询。这个一次性检索策略假设所有相关信息都可以提前收集。然而，对于涉及多跳推理或长格式生成的任务，模型的信息需求可能会在生成过程中演变。例如，中间推理步骤可能会引入新的子问题或需要澄清先前未见过的概念。为了支持这种场景，检索必须与生成动态集成，使模型能够在生成

[1]主要联系人: swh22@mails.tsinghua.edu.cn
[2]电子邮件：aiqy@tsinghua.edu.cn
[3]电子邮件：jingtaozhan@gmail.com

过程中访问信息。由于标准 RAG 在生成之前进行检索，它通常无法在生成过程中提供 LLM 所需的所有知识。

标准 RAG 的另一个基本限制在于检索到的知识如何被整合到模型中。具体来说，标准 RAG 采用上下文内知识注入的方法，即将检索到的文档附加到输入提示中，并与查询一起处理。虽然这种方法实现起来相对简单，但它带来了几个关键的挑战。首先，当相关知识的量很大时，直接将其附加到输入中会显著增加上下文长度，从而在推理过程中导致更高的计算成本和延迟。此外，随着上下文的加长，模型的注意力变得更加分散，降低了其对最相关信息的关注能力 [? ?] 。更根本的是，LLM 使用上下文信息的方式与其利用内部参数知识的方式本质上是不同的。经验研究表明，LLM 将其大部分事实知识编码在其神经架构的参数中，特别是在前馈网络层中 [? ? ?] 。相反，输入上下文中提供的信息仅通过自注意机制中的键值对的动态计算来影响模型，而不会被整合到模型的内部参数知识表示中。这种内部和上下文知识处理方式的根本差异导致对外部知识的使用较弱且不太可靠：LLM 可能无法在输出中始终如一地依据检索到的内容，特别是当这些知识与其内部参数知识冲突时。

这些限制激发了更先进的 RAG 范式的发展，旨在克服静态检索的约束和上下文知识注入的局限性。最近有两个代表性的方向受到显著关注：动态 RAG 和参数 RAG。动态 RAG 使检索能够与生成动态集成，允许系统根据不断变化的上下文按需访问相关信息，这对多跳推理和复杂生成任务特别有利。从另一个角度看，参数 RAG 旨在将外部知识直接注入模型的参数中，实现更深层次的整合，并允许 LLM 以利用其内部参数知识的方式使用外部知识。

本教程对这些新兴的 RAG 范式进行了全面介绍，为参与者提供了构建新一代检索增强系统的理论见解和实用技术。

2.2 动态检索增强生成

Dynamic RAG 是一种新兴的范式，克服了标准 RAG 的局限性。与依赖初始查询触发的单一检索步骤的标准 RAG 不同，Dynamic RAG 支持多轮检索，以适应 LLM 在生成过程中的不断变化的信息需求 [? ? ? ? ?] 。在每个检索步骤中，何时检索和检索什么的决定可以由 LLM 本身做出（例如，通过生成指示需要外部知识的特殊标记 [?] ），或者由一个外部系统做出，该系统监控模型的生成状态以检测不确定性 [? ?] 。同样，检索查询可以直接由 LLM 制定，或通过一个外部查询生成模块制定。通过在生成过程中相关点迭代整合检索到的外部知识，Dynamic RAG 使模型能够生成更准确、上下文感知和全面的响应，显著提升在复杂任务上的表现，如多跳推理和长篇文本生成。

最近的进展已证明这一范式的有效性。例如，Self-Reflective RAG (Self-RAG) [?] 引入了“反思标记”作为显示模型需要外部知识来辅助生成的显性信号。这些标记触发系统即时检索，以便大型语言模型（LLM）可以根据需要主动检索外部信息。实验结果表明，这种策略在需要精确事实支持的任务上优于静态 RAG。FLARE [?] 引入了一种不确定性感知检索机制，该机制在解码期间监控标记级生成概率。当检测到低置信度标记时，系统会自动生成检索查询，并用检索到的信息更新生成上下文。这一机制使模型能够纠正潜在的幻觉并提高事实准确性。DRAGIN [?] 通过提出一个轻量级 RAG 框架改进了 FLARE，该框架动态确定生成期间 LLM 的信息需求何时以及检索何种信息。与 Self-RAG 和 FLARE 不同，DRAGIN 不需要额外的训练或提示修改。它通过分析模型的内部信号，如注意力分布和预测熵，来建模检索需求，并使用自注意力根据当前上下文制定精确的检索查询。实验结果表明，DRAGIN 在需要动态访问外部知识的复杂任务上表现优异。SeaKR [?] 引入了一种自知的检索机制，利用 LLM 的内部状态来量化标记级不确定性。当不确定性超过预定阈值时，系统触发外部检索，并根据减少不确定性的潜力对检索到的段落进行重新排序。这种有针对性的检索方法提高了事实准确性，同时避免了不必要的开销。

总之，动态 RAG 通过支持模型以按需、迭代的方式获取信息，提供了一种比标准 RAG 更灵活的替代方案。通过自我反思、不确定性检测和令牌级影响分析等机制，动态 RAG 允许模型自适应地识别何时需要检索以及在触发检索模块时需要检索什么信息。这种生成与检索之间的动态互动不仅提高了事实一致性，还增强了模型处理复杂的多步骤推理任务的能力，而这些任务无法通过静态的一次性检索有效支持。

2.3 参数化检索增强生成

虽然动态 RAG 解决了在生成过程中检索 当 和 什么 的问题，但参数化检索增强生成（Parametric RAG）关注于一个更为根本的挑战：如何将检索到的知识整合到模型中。尽管 RAG 文献中相关研究不断增多，大多数现有方法采用上下文知识注入，即将外部段落附加到输入提示中，并与查询一起处理。然而，如 § 至 2.1 中讨论的，这种方法在效率和效果上存在显著的局限性 [? ? ?] 。为克服这些局限，参数化 RAG 引入了一种替代范式，直接将检索到的知识整合到模型的参数中。参数化 RAG 的核心挑战是设计一种转化，将文档 D 映射成一个插件参数模块 P ，以便在将 P 插入 LLM 后，模型能够内化嵌入在 D 中的知识。为解决这一挑战，已有两种主流方法被提出用于实现文档到参数的转化。

第一类方法采用基于训练的方法来获取特定文档的参数化表示。对于每个文档，通过扩展原始内容来构建一个合成数据集，通常通过重写或生成问答对来实现。然后，在这个数据集上微调轻量级适配器模块，例如 LoRA [?] 。结果生成的 LoRA 模块作用为插件参数，编码文档的知识，并作为其参数化表示。这一类别的一个代表性方法是 PRAG [?] ，其采用两阶段的流程：在离线阶段，每个文档通过上述训练过程编码为一个特定于文档的 LoRA 模块；在推理时，检索出最相关的 k 模块并将其合并到基础模型中，使得更新后的大型语言模型能够进行特定查询的生成。

第二类方法通过利用在线参数生成策略，消除了对文档特定的微调或存储的需求。通过训练一个小型超网络来将每个文档的语义嵌入映射到一个参数高效的模块，而不是对单独的适配器进行预训练。在推理阶段，检索到的文档的语义嵌入通过这个超网络来动态生成即时的插件参数。该方法实现了实时适应，并显著降低了维护大量参数模块库的成本。DyPRAG [?] 通过引入一个动态参数翻译器，在即时生成基于检索文档的语义嵌入的文档特定参数表示方面，体现了这一方法。该设计使得参数化 RAG 在不维护大量参数模块库的情况下实现了可扩展性和效率。

总之，参数化 RAG 代表从上下文到参数级知识注入的范式转变，为将外部信息整合到 LLM 中提供了更为有效的解决方案。

3 目标

本教程旨在为与会者提供对检索增强生成（RAG）最近进展的全面理解，特别是关注于两个新兴且互补的范式：动态 RAG 和参数 RAG。通过参加本教程，参与者将：

- 清晰理解 RAG 的基础知识，包括其当前发展状况和标准 RAG 框架的主要局限性。
- 了解动态 RAG 的动机、架构和代表性技术，该方法将检索与生成交错进行，以支持对外部知识的实时、适应性访问。
- 了解参数化 RAG 的原理和实现策略，这是一种新兴的 RAG 范式，可以将外部知识直接注入到模型参数中。
- 深入了解不断发展的 RAG 领域中出现的研究趋势、开放性挑战和未来方向。

4 与信息检索社区的相关性

检索增强生成（RAG）通过结合传统检索技术与大型语言模型（LLM）的强大语言生成能力，直接应对现代信息检索的关键挑战。与主要关注于对单一查询作出响应而检索静态文档的传统信息检索方法不同，近年来许多新兴的 RAG 框架引入了检索与生成之间的动态交互，从而使系统能够更有效地解决复杂查询并支持多跳推理。本教程针对检索和生成交汇处的核心挑战，特别强调两个新兴且互补的研究方向：动态 RAG 和参数化 RAG。动态 RAG 通过在生成过程中交错检索来研究何时以及检索什么，从而解决一次性检索策略的局限性。参数化 RAG 则探索如何将检索到的知识编码到模型的参数中，从而实现更有效的知识注入。这两个方向都代表了信息检索领域的关键进展，因为它们为构建更具适应性、准确性和效率的检索增强生成系统提供了新颖的解决方案。

虽然在 SIGIR 2022 [**?**] 和 ACL 2023 [**?**] 的之前教程中已经介绍了标准 RAG 方法的进展，但它们没有讨论基于动态检索或参数知识注入的新兴 RAG 范式。本教程通过提供这些两种互补 RAG 范式的实用见解和技术深度，填补了这一空白，为信息检索研究人员开发下一代 IR 系统提供了见解。

5 格式和时间表

这是一个半天（3 小时）的讲座式教程，其中包括安排的休息时间。它将在现场进行，至少两位讲师计划亲自参加会议。

- 介绍和背景（30 分钟）：
  - 教程概览
  - RAG 的基础
  - 新的 RAG 范式
  - 标准 RAG 的局限性
- 主题 1 —动态 RAG (50 分钟):
  - 动机：为什么需要动态 RAG?
  - 何时检索：在生成过程中检索模块应何时被触发?
  - 检索什么：我们如何制定能够反映模型实时信息需求的查询?
  - 开放性挑战和未来方向
- Q & A (10 分钟)
- 休息（30 分钟）
- 主题 2 —参数化 RAG (40 分钟)：
- 总结与未来方向（10 分钟）：总结教程的主要见解，讨论开放问题，以及在检索增强生成领域的未来研究机会。
- Q & A（10 分钟）

6 支持材料

本教程的补充材料将包括一份描述技术内容细节的手稿，以及按主题分组的重要参考文献和相关教程的综合列表。教程幻灯片也将提供。此外，我们将维护一个 GitHub 库，调查和分类与该主题相关的所有论文，以促进进一步阅读和探索。所有材料将通过专门的网站提供访问权限，该网站将在教程前与参与者分享。