ARES：一种用于检索增强生成系统的自动化评估框架

作者信息：

- Jon Saad-Falcon, 斯坦福大学 (jonsaadfalcon@stanford.edu)
- Omar Khattab, 斯坦福大学 (okhattab@stanford.edu)
- Christopher Potts, 斯坦福大学 (cgpotts@stanford.edu)
- Matei Zaharia, Databricks and 加州大学伯克利分校 (matei@databricks.com)

摘要

评估检索增强生成（RAG）系统传统上依赖于对输入查询、待检索段落和待生成响应的手动标注。我们引入了 ARES，一个**自动化 RAG 评估系统**，用于从上下文相关性、答案忠实度和答案相关性等维度评估 RAG 系统。通过创建自身的合成训练数据，ARES 能够微调轻量级语言模型（LM）“裁判”，以评估各个 RAG 组件的质量。为减轻潜在的预测错误，ARES 利用一小部分人工标注的数据点进行预测驱动的推断（PPI）。在 KILT、SuperGLUE 和 AIS 中的八个不同知识密集型任务中，ARES 在评估过程中仅使用数百个人工标注，便能准确地评估 RAG 系统。此外，即使在被评估的 RAG 系统中所使用的查询和/或文档类型发生变化，ARES 的“裁判”模型在领域迁移后依然保持有效和准确。我们的代码和数据集已在 Github 上公开。

1. 引言

对于不同的数据领域、语料库大小以及成本/延迟预算，RAG 系统的最佳设计并非普遍适用。为了调整自己的 RAG 系统，从业者传统上需要针对其目标领域手动标注测试问题、待检索的段落（用于评估检索器）以及待生成的响应。或者，他们也可以通过收集用户偏好来在生产环境中评估不同方案，以比较候选系统。然而，这两种策略都要求高度的专业知识，并带来相当大的标注成本。

基于模型的评估是一种测试生成内容质量的低成本策略。例如，开源的 RAGAS 框架利用大语言模型（LM）提示来评估检索信息的**相关性**以及生成响应的**忠实度**和**准确性**。然而，这类策略目前依赖于一组固定的、启发式手写的提示进行评估，对各种评估环境的适应性很小，且无法保证质量。

为了快速准确地评估 RAG 系统，我们提出了 ARES，即**自动化 RAG 评估系统**。ARES 是首个为 RAG 流程的每个组件生成定制化 LLM“裁判”的自动化评估系统，与 RAGAS 等现有方法相比，在评估精度和准确性上取得了显著提升。此外，与现有的 RAG 评估系统不同，ARES 利用预测驱动的推断（PPI）为其评分提供置信区间。给定一个文档语料库和一个 RAG 系统，ARES 会报告三个评估分数：**上下文相关性**（检索到的信息是否与测试问题相关）、**答案忠实度**（语言模型生成的响应是否恰当地基于检索到的上下文）以及**答案相关性**（响应是否也与问题相关）。一个好的

RAG 系统能够找到相关的上下文，并生成既忠实又相关的答案。

许多现有的 RAG 评估框架需要大量的人工标注来进行评分。ARES 通过仅需三个输入显著提高了评估时的数据效率：一个**领域内段落集**，一个约 150 个或更多已标注数据点的**人工偏好验证集**，以及少量领域内查询和答案的**少样本示例**（例如五个或更多示例），这些示例用于在合成数据生成中提示 LLM。

给定领域内段落的语料库，ARES 分三个阶段进行。首先，它利用一个 LM 从语料库的段落中构建一个**合成的问答对数据集**。其次，它定义了三个独立的"裁判"模型来执行三个分类任务（上下文相关性、答案忠实度和答案相关性）。这些"裁判"是针对对比学习目标进行微调的轻量级模型。第三，ARES 使用**预测驱动的推断（PPI）对被评估的不同 RAG 系统进行评分，以提高基于模型的评估准确性，并为 RAG 评分提供统计置信区间。PPI 利用一小部分人工标注的数据点来计算其置信区间；我们将这个标注集指定为我们的人工偏好验证集**，它由大约 150 个或更多的标注数据点组成，这些数据点为上下文相关性、答案忠实度和答案相关性指定了正面和负面的例子。

我们进行了广泛的实证评估，证明 ARES 在 KILT 和 SuperGLUE 的六个知识密集型数据集上准确地对 RAG 系统进行评分，在上下文相关性和答案相关性评估准确性上，平均分别比现有的自动化评估方法（如 RAGAS）高出 59.3 和 14.4 个百分点。此外，ARES 在 AIS 归因数据集中准确计算了答案幻觉的出现次数，预测的答案幻觉平均值与真实值的差距在 2.5 个百分点以内。

与基于标注的评估方法相比，ARES 的准确性和效率显著更高，比基线方法所需的标注量减少了 78%。我们还发现，ARES 能够持续地区分那些在真实指标上仅相差几个百分点的有竞争力的 RAG 系统。这种精确性使 ARES 能够指导有竞争力的方案和配置的开发与比较。

我们的 ARES 代码和数据集已在 Github 上公开。

2. 相关工作

检索增强生成（RAG）现在是通过将大语言模型（LLMs）与检索系统相结合来增强其能力的常用策略。通过检索，RAG 帮助 LM 系统收集特定领域的知识，将生成内容基于事实信息，并通过引用来源提供一定程度的透明度或可解释性。

多种基于 LLM 的评估技术已经出现，用于衡量 LLM 系统的性能。这对于在难以从头开始构建传统基准数据集的新环境中快速部署至关重要。早期的尝试是直接使用开箱即用的 LLM，如 MT-Bench 和 Chatbot Arena。AutoCalibrate 试图将 LLM"裁判"与人类偏好对齐，利用自我修正的提示来迭代改进 LLM"裁判"。然而，AutoCalibrate 对其预测的准确性不提供任何统计保证。其他工作已使用 LLM 提示来评估自然语言生成任务中的系统质量，例如翻译、摘要和对话。

在知识密集型 NLP 任务的背景下，LLM 已被探索用于评估 LLM 的归因和事实性。新的指南如 LongEval 和数据集如 Hagrid 和 ALCE 为分析知识密集型 LLM 流程提供了资源。

与 ARES 最相关的两个项目是 EXAM 和 RAGAS。为了评估 RAG 系统，EXAM 度量标准估计一个读者（模拟为问答系统）能根据生成的回答正确回答多少考试问题。这需要一组查询，每个查询都附有几个相关的子问题，这增加了 ARES 所没有的负担。RAGAS 基于少数启发式手写的提示。这些提示对新的 RAG 评估环境（例如，新的语料库）的适应性很小，并且正如我们在评估中所示，其性能远不如 ARES。

3. ARES 框架

ARES 分三个阶段进行（见图 1）。需要三个输入：一个**领域内段落集**，一个约 150 个或更多已标注数据点的**人工偏好验证集**，以及少量领域内查询和答案的**少样本示例**（五个或更多示例），这些示例用于在合成数据生成中提示 LLM。准备好输入后，我们首先从目标语料库的段落中生成合成的查询（及其答案）。然后，我们使用这些查询-段落-答案三元组来训练 LLM“裁判”。随后，我们将这些“裁判”应用于任何 RAG 系统，对其领域内查询-文档-答案三元组的样本进行评分，并使用我们的**人工偏好验证集**通过预测驱动的推断（PPI）来估计每个 RAG 系统质量的置信区间。

3.1 LLM 生成合成数据集

我们使用生成式 LLM 从语料库段落中生成合成的查询和答案。生成的数据代表了查询-段落-答案三元组的正面和负面示例（例如，相关/不相关的段落和正确/不正确的答案）。在生成过程中，LLM 使用我们输入的少样本示例集，其中领域内段落映射到领域内查询和答案；然后模型从给定的领域内段落生成一个合成的问题和答案，使我们能够创建正面和负面的训练示例。

为了创建我们的合成数据，我们主要使用 FLAN-T5 XXL。ARES 与这个模型配合得很好，但我们的系统最终也可以使用其他高质量模型来生成合成的查询和答案。然后，我们通过测试给定查询是否能使用其检索器将其原始段落作为首要结果检索出来，来过滤掉低质量的查询。

为了生成用于微调我们 LLM“裁判”的负样本，我们依赖两种新颖的策略，每种策略生成相同数量的负样本：

- **弱负样本生成**：对于上下文相关性的负样本，我们随机抽样与给定合成查询无关的领域内段落。对于答案忠实度和答案相关性的负样本，我们从其他段落中随机抽样合成生成的答案。
- **强负样本生成**：对于上下文相关性的负样本，我们从与黄金段落相同的文档中随机抽样领域内段落。对于答案忠实度和答案相关性的负样本，我们提示 FLAN-T5 XXL 生成一个矛盾的答案。

总的来说，为评估上下文相关性和答案相关性生成的负样本数量等于生成的正样本数量。

3.2 准备 LLM“裁判”

为了准备我们的 RAG 评估“裁判”，我们使用我们的合成数据集来微调 DeBERTa-v3-Large“裁判”，以评估三种不同的能力：

- **上下文相关性**：返回的段落是否与回答给定查询相关？

- **答案忠实度**：生成的答案是否忠实于检索到的段落，还是包含了超出段落范围的幻觉或推断性陈述？
- **答案相关性**：生成的答案在给定查询和检索段落的情况下是否相关？

对于每个指标，一个带有二元分类器头的独立 LLM 被微调以分类正面和负面示例。对于每个连接的查询-文档-答案，一个 LLM“裁判”必须为该“裁判”的指标将三元组分类为正面或负面。

3.3 使用置信区间对 RAG 系统进行排名

一旦我们准备好了 LLM“裁判”，我们需要用它们来对竞争的 RAG 系统进行评分和排名。为此，ARES 对每个 RAG 方法产生的领域内查询-文档-答案三元组进行抽样，然后由“裁判”对每个三元组进行标注，预测其上下文相关性、答案忠实度和答案相关性。通过平均每个领域内三元组的单个预测标签，我们计算出 RAG 系统在三个指标上的性能。

原则上，我们可以简单地将这些平均分数作为每个 RAG 系统的质量指标报告。然而，这些分数反映的是完全未标注的数据，预测来自一个合成训练的 LLM“裁判”，因此它们可能不完全准确。作为一种极端的替代方案，我们可以仅使用之前讨论的小型人工偏好验证集进行评估。然而，基于标注的评估方法需要对每个 RAG 系统分别标注更多的生成输出，这在时间和资金上都可能成本高昂。

为了结合两者的优点，从而提高评估的精确度，ARES 使用**预测驱动的推断（PPI）**来预测系统分数。PPI 是一种最新的统计方法，通过利用对更大规模未标注数据点的预测，为一小组标注数据点（即我们的验证集）提供更紧密的置信区间。PPI 可以利用标注数据点和 ARES“裁判”对未标注数据点的预测来构建我们 RAG 系统性能的置信区间。

通过用带有机器学习预测的更大规模数据点集来增强人工偏好验证集，PPI 可以为机器学习模型性能开发出可靠的置信区间，这些置信区间优于以前的经典推断方法。

4. 实验

4.1 模型

对于我们的微调“裁判”，ARES 依赖于使用 LLM 生成廉价但高质量的合成查询和答案。我们使用 **FLAN-T5 XXL** 生成合成数据集，并选择 **DeBERTa-v3-Large** 作为微调的 LLM“裁判”。对于我们的上下文学习基线，我们使用 OpenAI 的 **gpt-3.5-turbo-16k**。

4.2 数据集

我们的核心实验目标是全面展示 ARES 可以有效应用的场景。为了测试跨多种类型的查询、文档和答案，我们从广泛使用的 KILT 和 SuperGLUE 基准测试中选择了所有适合 RAG 的数据集。

- **KILT**: Natural Questions (NQ), HotpotQA, FEVER, 和 Wizards of Wikipedia (WoW)。
- **SuperGLUE**: MultiRC 和 ReCoRD。

我们通过人工创建模拟的 RAG 系统来测试 ARES，这些系统在我们的评估指标上具有已知的准确

率，范围从 70%到 90%不等，间隔为 2.5%。

4.3 指标

我们使用**肯德尔等级相关系数（Kendall's τ）**来计算正确排名与 ARES 排名之间的相关性。肯德尔 τ 是衡量成对比较准确性的流行指标，非常适合评估排名系统。

5. 结果与分析

5.1 ARES 排名

在几乎所有来自 KILT 和 SuperGLUE 数据集的设置中，ARES 提供的 RAG 系统排名都比 RAGAS 更准确。ARES 的肯德尔 τ 在**上下文相关性**上平均高出 **0.065**，在**答案相关性**上平均高出 **0.132**。与 RAGAS 相比，ARES 的 LLM“裁判”在预测上下文相关性和答案相关性方面也明显更准确，准确率分别高出 **59.9** 和 **14.4** 个百分点。

与基于抽样标注的方法相比，ARES 在准确性上更高，同时使用的标注量减少了 78%，显示出更高的数据效率。与 GPT-3.5“裁判”相比，ARES 也提供了更准确的排名。

5.2 ARES 在 AIS 上的表现

为了评估 ARES 是否能有效衡量真实 RAG 系统中的答案忠实度，我们在 AIS 归因基准上测试了 ARES。结果表明，ARES 能够有效评分 AIS 数据集，与正确分数的差距在 2.5 个准确率点以内，证明了其区分忠实答案和幻觉答案的能力。

5.3 ARES 对现有 RAG 系统的排名

我们还评估了 ARES 是否能对现有的 RAG 系统进行评分和排名。结果发现，ARES 能够可靠地对真实世界的 RAG 系统进行评分和排名，上下文相关性的肯德尔 τ 平均为 0.91，答案相关性为 0.97，显著优于 RAGAS。

5.4 跨领域应用的优势与局限

ARES 中使用的 LLM“裁判”在跨领域应用中表现出强大的泛化能力，即使在查询类型、文档类型或两者都发生变化时也是如此。然而，当领域发生更剧烈的变化时，例如语言切换（如英语到西班牙语）、从文本到代码的转换，或从检索文本到提取实体，LLM“裁判”的泛化能力会下降。

6. 结论

我们介绍了 ARES，一个新颖的 RAG 自动化评估框架。ARES 为微调轻量级 LLM“裁判”提供了一个新颖的训练流程，并且只需要最少的人工标注。实验证明，ARES 在上下文相关性、答案忠实度和答案相关性方面能够准确地评分和排名 RAG 系统，优于现有的 RAGAS 框架。

7. 局限性

ARES 依赖于一小组人工标注，这在专业领域可能需要具备专业知识的标注员。此外，ARES 中使用的 LLM 需要具有大量存储空间的 GPU 硬件，这可能不是所有研究人员和从业者都能轻易获得的。最后，我们的评估主要集中在英语上，未来的工作应该探索 ARES 在其他语言中的应用。