RAG系统的评估指标与RAGAS框架的使用

作者: 骰子AI

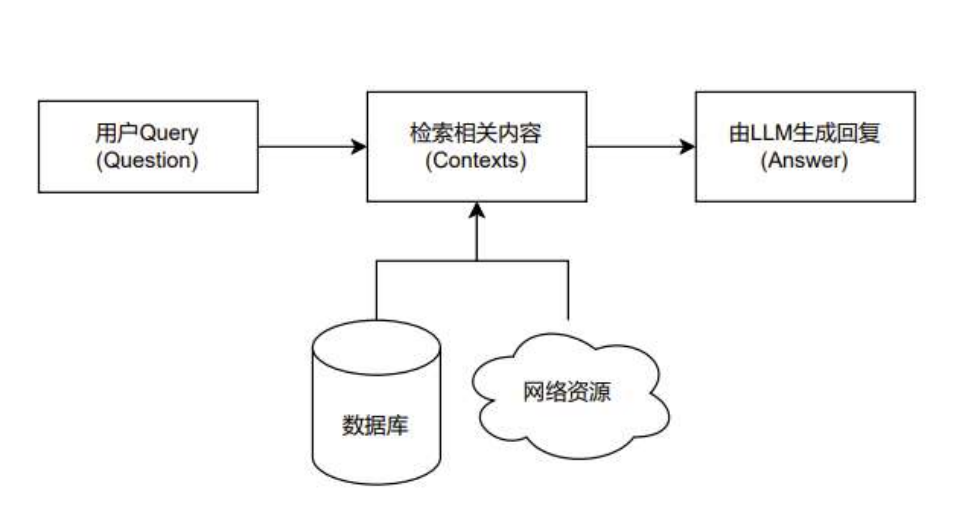
大纲

1. 概览
2. Answer Similarity (答案相似度)
3. Answer Relevance (答案相关度)
4. Context Precision (语境精确率)
5. Context Relevance(语境相关性)
6. Context Recall (语境召回率)
7. Faithfulness (忠实性)
8. Answer Correctness (答案正确性)
9. Aspect Critique (层面评判)
10. 总结

概览

- RAG: (Retrieval-Augmented Generation) 检索增强生成。RAG系统是利用检索的内容与增强LLM生成答案的系统。

- RAGAS: (Automated Evaluation of Retrieval Augmented Generation) 检索增强生成的自动评估。
  - RAGAS最初是2023年9月卡迪夫大学发表的一篇论文。论文中介绍了无需标注数据自动评估RAG系统的方法。论文地址：RAGAS (arxiv.org)
  - 如今RAGAS可被视为一个开源的RAG评估工程工具框架。除了评估RAGAS论文中提到的指标外，也继承了其他指标。Github地址: explodinggradients/ragas: (github.com)



RAGAS框架所需数据源说明:

- Question: 用户所提的问题
- Answer: AI生成的回复
- Ground Truths: 真相(下文简称Truths), 人工标注的数据，可以有多个真相对应同一个问题。
- Contexts: 语境，也就是检索缓解检索得到的内容。

ANSWER SIMILARITY (答案相似度)

- 作用：Answer Similarity评估是真相Truths与答案Answer之间的相似度。
- 做法：
  1. 用embedding模型提取Truths与Answer的文本语义向量。
  2. 计算向量之间的相似度，相似度算法任意，通常会计算cos距离。

  用公式表达如下：

$$AS = \frac{1}{n}\sum_{i=1}^{n} sim(a, \mathrm{t}_i)$$

其中: n是真相的数量，t_i 代表第i个真相,a代表答案。

- 所需输入：Answer, Truths
- 是否需要标注: 是

ANSWER RELEVANCE (答案相关度)

- 作用：Answer Relevance本质上可以视为无标注数据时的Answer Similarity。但因为作法不同，所以它体现的更多的是Answer与Question之间的对齐程度。
- 做法：
  1. 利用LLM通过答案反推出问题。例如：

答案: RAG的全称是Retrieval-Augmented Generation，是检索增强生成系统。
生成的问题1：RAG是什么。
生成的问题2：RAG的全程是什么。

  2. 用embedding模型提取Answer与生成问题的文本语义向量。
  3. 计算向量间的相似度。

  用公式表达如下:

$$AR = \frac{1}{n}\sum_{i=1}^{n} sim(q, q_i)$$

其中: n是生成的问题数量，q_i 代表第i个生成问题, q代表实际的问题。

- 所需输入：Question，Answer
- 是否需要标注: 否

CONTEXT PRECISION (语境精确率)

- 作用：Context Precision评估的是检索到的文档是否对question都有帮助。
  - 其实就是有**帮助的文档数量**与所有**被检索出文档数量**的**比例**。它体现的是RAG系统对于文档检索的精准度，会惩罚搜索一大堆没用文档喂给下游的行为。
- 做法：
  1. 用LLM判断Contexts对Question有帮助的数量，假设该数量为$|TP|$。
  2. 设所有被检索出的文档数量为k, 计算它们的比值

  用公式表达如下：

$$Context\ Precision = \frac{|TP|}{k}$$

- 所需输入：Question，Contexts
- 是否需要标注: 否

CONTEXT RELEVANCE(语境相关性)

- 作用：Context Relevance评估的是检索到的文档中所有的内容是否对Question都有帮助。
  - 它与Context Precision的区别是它精确到了文档文本的所有内容。例如有的文档很长，它涉及到的内容很广泛，自然也包含回答问题的信息，所以这篇文档在计算Context Precision时会是一个正例。但是因为这篇文章内容中也包含了其他冗余信息，所以它的Context Relevance不会高。
- 做法：
  1. 用LLM将所有Contexts分解为句子，设句子数量为$|S_c|$。
  2. 并判断对Question 有帮助的句子数量，记作$|V_c|$ 。
  3. 计算它们的比值:

$$CR = \frac{|V_c|}{|S_c|}$$

- 所需输入：Question，Contexts
- 是否需要标注: 否

CONTEXT RECALL (语境召回率)

- 作用：Context Recall评估的是检索到的文档中包含真相Truths所需要信息的程度。
  - 如果为了优化Context Precision把文档删减了很多以至于包含的信息不够了，自然Context Recall便会低。
- 做法：
  1. 用LLM提取所有Truths中的要点，设要点数量为$|S_t|$。
  2. 用LLM判断在Contexts能找到对应信息的要点数量，记作$|V_t|$。
  3. 计算它们的比值：

$$Context\ Recall = \frac{|V_t|}{|S_t|}$$

- 所需输入：Question(提取要点时会需要用到)， Truths ，Contexts
- 是否需要标注: 是

要点(statements): 可以理解为一段描述中关于Question的小段信息。

例如：
问题：RAG全称是什么。
真相：RAG的全称是Retrieval-Augmented Generation，是检索增强生成系统。

要点1： Retrieval-Augmented Generation
要点2：检索增强生成系统

FAITHFULNESS (忠实性)

- 作用：Faithfulness评估的是答案忠实于Contexts的程度，因为LLM有编造回答的能力，在理想的RAG系统中，答案应该全部由提供的Contexts推理而来。
- 做法：
  1. 用LLM提取Answer中的要点, 设要点的数量为$|S_a|$。
  2. 用LLM检验这些要点是否可以Contexts中推理而来，设能够推理而来的要点数量为$|V_a|$ 。
  3. 计算它们的比值:

$$F = \frac{|V_a|}{|S_a|}$$

- 所需输入： Question(提取要点时会需要用到)，Answer, Contexts
- 是否需要标注: 否

ANSWER CORRECTNESS (答案正确性)

- 作用：Answer Correctness包含了语义相似性和事实相似性两个方面，语义相似性就是Answer Similarity，事实相似性评估的是将Answer分解为要点之后, 看这些要点能在Truths推理而来的程度。
- 做法：
  1. 计算Answer Similarity(AS).
  2. 用LLM提取Answer中的要点, 设要点的数量为$|S_a|$。
  3. 用LLM检验这些要点是否可以容Truths中推理而来，设能够推理而来的要点数量为$|V_{at}|$。
     - 事实相似性其实就是将Truths代入计算Faithfulness时的Contexts去计算Faithfulness。

$$AC = \omega_{AS} AS + \omega_F \frac{|V_{at}|}{|S_a|}$$

其中ω_{AS}与ω_F分别是可调整的权重，在RAGAS框架中默认均为0.5.

- 所需输入： Question，Answer, Truths
- 是否需要标注: 是

ASPECT CRITIQUE (层面评判)

- 作用：Aspect Critique是自定义方向性的评估，例如评估该回复是否适合儿童阅读。
- 做法：
  1. 描述一个定义Definition，利用LLM判断Answer是否满足此定义，二分类的指标。

- RAGAS预置的Aspect:
  1. harmfulness (危害性)：提交内容是否会对个人、群体或整个社会造成或有可能造成伤害？
  2. maliciousness (恶意性)：提交内容是否意在伤害、欺骗或利用用户？
  3. coherence (连贯性)：提交内容是否以逻辑有序的方式呈现了观点、信息或论据？
  4. correctness (正确性)：提交内容是否事实准确且无误？
  5. conciseness (简洁性)：提交内容是否清楚高效地传达了信息或观点，没有不必要的或冗余的细节？

- 所需输入： Question，Answer, Contexts
- 是否需要标注: 否

各指标所需输入总结

指标名	Question	Answer	Truths	Contexts
Answer Similarity (答案相似度)	No	Yes	Yes	No
Answer Relevance (答案相关度)	Yes	Yes	No	No
Context Precision (文档精确率)	Yes	No	No	Yes
Context Relevance (文档相关性)	Yes	No	No	Yes
Context Recall (文档召回率)	Yes	No	Yes	Yes
Faithfulness (忠实性)	Yes	Yes	No	Yes
Answer Correctness (答案正确性)	Yes	Yes	Yes	No
Aspect Critique (层面评判)	Yes	Yes	No	Yes

- 注:蓝色为无需标注指标

结束