### **3.1 评估指标**

为了评估 HybridRAG 的效果，作者建立一整套评估指标体系：

#### **3.1.1 忠实度**

忠实度是一个关键的评估指标，用于衡量生成的答案能在多大程度上从提供的上下文中推断出来。对忠实度指标的实现包括一个两步流程：

• ****语句提取****：利用 LLM 将生成的答案分解为一系列简洁的陈述。这一步骤对于将复杂答案拆解为更易于管理和验证的单元至关重要。用于此步骤的提示是：

“给定一个问题和答案，请从给定答案的每个句子中创建一个或多个陈述。

问题：[问题]

答案：[答案]”。

• ****语句验证****：对于每个提取的陈述，使用 LLM 来确定它是否可以从给定的上下文中推断出来。这个验证过程使用的提示是：

“考虑给定的上下文和以下陈述，然后判断它们是否得到了上下文中信息的支持。在得出每个陈述的结论（是/否）之前，为每个陈述提供简短的解释。在给定格式的末尾，按顺序为每个陈述提供最终的结论。不要偏离指定的格式。”

声明：[声明 1]

……

声明：[声明 n] 。

忠实性得分计算方法为 F=V/S，这里 𝑉 代表得到支持的语句数目，而 𝑆 代表语句的总数目。

#### **3.1.2 答案相关性**

答案相关性指标评估生成的答案在多大程度上解决了原始问题，不考虑事实准确性。该指标有助于识别不完整的答案或包含无关信息的回复。

包含以下步骤：

• 问题生成：我们促使大型语言模型（LLM）根据给定的答案生成 n 个潜在问题：

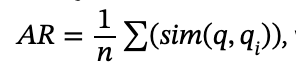
“为给定的答案生成一个问题。

答案：[答案]” 。

• 然后，使用 OpenAI 的 text-embedding-ada-002 模型获取所有生成问题和原始问题的嵌入。

• 接着，我们计算每个生成问题的嵌入与原始问题的嵌入之间的余弦相似度。

• 最后，答案相关性得分（AR）计算为所有生成问题的平均相似度。这里的 sim(𝑞, 𝑞𝑖) 表示原始问题 𝑞 与每个生成问题 𝑞𝑖 嵌入向量之间的余弦相似度。



#### **3.1.3. 上下文精度**

一种衡量检索上下文片段与特定问题的真实情况相关性的指标。度量在上下文的前 K 个排名中，相关项所占的比例。上下文精度的计算公式是在 K 的排名中，每个排名 k 的精度与一个二元相关性指示器 𝑣𝑘 的乘积之和，然后除以上 K 个结果中相关项的总数量。

每个排名 k 的精度是通过 k 处的真阳性数与 k 处的真阳性数和假阳性数的总和之比来确定的。

这一指标有助于评估上下文对真实情况的支撑力度，更高的分数表示更精确的匹配度。

#### **3.1.4. 上下文召回率**

该指标用于衡量检索到的上下文与标准答案（作为确切的正确回答）的契合度。

通过对比标准答案中的每一句话，判断其是否能在检索到的上下文中找到依据，来量化这一指标。

上下文召回率的计算公式是，可追溯至上下文的标准答案句子数量与标准答案中句子总数的比值。

数值越高，介于 0 到 1 之间，表示上下文与答案的契合度越好，即上下文召回率越高。