使用 DeepEval 指標，優化 RAG 系統

一、 系統說明與實驗設定

本作業以「台灣自來水公司常見問答」為主題，建立一個模擬的 RAG 評估流程，並使用 DeepEval 提供的五項指標來分析系統回答品質。

透過固定的通用 DEFAULT_CONTEXT，模擬「RAG 系統已取得背景知識，但不一定完全對應問題」的情境，藉此觀察 DeepEval 各指標的行為差異。

使用模型:

- LLM：gemma-3-27b-it
- Temperature = 0.1~0.5

評估資料:

- 從題庫中隨機抽樣 5 筆
- 問題型態：口語化、生活情境導向（水質、水費、帳單）

評估輸入：

- input：使用者問題
- actual_output：LLM 生成答案
- expected_output：未特別指定（部分指標不使用）
- retrieval_context：通用水務背景說明（DEFAULT_CONTEXT）

二、DeepEval 實際評估結果

q_id	Faithfulness	Answer Relevancy	Contextual Recall	Contextual Precision	Contextual Relevancy
25	1.000	0.944	0.375	1.000	0.000
7	1.000	1.000	0.000	1.000	0.000
5	0.929	0.944	0.571	1.000	0.333
9	1.000	1.000	0.550	0.000	0.000
27	1.000	0.762	0.000	0.000	0.000

三、DeepEval 五項指標說明與結果分析

（1）Faithfulness（忠實度）

平均表現：接近 1.0（極高）

Faithfulness 用來衡量「答案是否忠於 context」，是否產生幻覺。

本實驗觀察：

幾乎所有題目皆為 1.0

表示 LLM 的回答內容並未捏造 context 中不存在的資訊

結論：在回答水質與水費常見問題時，模型具有良好穩定性，不易產生幻覺。

（2）Answer Relevancy（答案相關性）

平均表現：約 0.9（良好）

此指標衡量「回答是否真的在回答問題本身」。

觀察：

- 大多數題目分數高於 0.9
- 僅在回答過於延伸或補充過多背景時略微下降（如 q_id 27）

✅ 結論：模型能正確理解問題，但可進一步要求「精簡回答」來提升此指標。

（3）Contextual Recall（上下文召回率）

整體表現：偏低，甚至為 0

Contextual Recall 的定義是：

「回答所需的關鍵資訊，是否存在於 retrieval_context 中」

📌 本實驗情境：

- 使用的是「通用 DEFAULT_CONTEXT」

```
DEFAULT_CONTEXT = [
    "自來水公司依照國家飲用水水質標準進行淨水與消毒處理，以確保供水安全與品質。",
    "自來水相關業務包含水費帳單寄送、繳費方式、電子帳單申請及用水問題諮詢等服務。",
    "若民眾在用水、水質或帳單方面遇到疑問，可洽詢自來水公司客服或至營業所辦理。"
]
```

- 並未針對每個問題檢索專屬段落
- 因此多數答案所需的具體資訊 並不存在於 context 中

✅ 結論：
Contextual Recall = 0 不代表系統錯誤。

（4）Contextual Precision（上下文精確度）

結果呈現兩極化（0 或 1）

此指標衡量：

context 中是否「大多是有用資訊」

觀察：

- 當問題能被通用水務背景概括時 → 得分 1.0
- 當問題非常具體（帳單、費率）→ 得分 0

解釋：

- DEFAULT_CONTEXT 沒有「錯」
- 但對於具體情境來說，相關性不足

結論：此指標成功反映「通用 context 無法取代真正的檢索結果」。

（5）Contextual Relevancy（上下文相關性）

整體表現偏低（0 ~ 0.33）

此指標評估：

整體 context 與問題的語意相關程度

結果分析：

- 幾乎所有題目都接近 0
- 因為 DEFAULT_CONTEXT 內容偏「官方背景說明」
- 與實際問題（藥味、水垢、污水費）關聯度有限

結論：
此結果非常適合用來說明「為什麼 RAG 一定要做檢索」。

四、從 DeepEval 指標得到的 RAG 優化啟示

指標	實驗結果	可行優化方向
Faithfulness	高	維持低 temperature
Answer Relevancy	高	限制回答長度
Contextual Recall	低	加入真正的文件檢索
Contextual Precision	不穩定	Re-rank + Top-K 篩選
Contextual Relevancy	低	Query Rewrite + 精準 chunk

五、總結

本次作業透過 DeepEval 五項指標，成功驗證以下幾點：

1. LLM 本身具備良好回答能力（Faithfulness / Answer Relevancy 高）
2. 沒有檢索就不是真正的 RAG（Contextual Recall / Relevancy 低）
3. DeepEval 能精準指出系統瓶頸位置

這使得 DeepEval 不只是評分工具，而是 RAG 系統設計與優化的診斷儀表板。