# 使用 DeepEval 指標優化 RAG 系統

HW Day6 — 學號 14112320[illegible] — 台灣自來水公司 AI 客服助手

## 一、系統架構總覽

本系統為台灣自來水公司的 AI 客服助手，基於 RAG（Retrieval-Augmented Generation）架構，處理約 71K 字元的 QA 知識庫（192 個切塊），回答 30 題口語化的用戶問題。從中隨機抽樣 5 題進行 DeepEval 完整評估。

### 系統流程

口語問題 → Query ReWrite → Hybrid Search（Dense + BM25）→ RRF 融合 → ReRank（LLM 評分）→ Top-3 段落 → LLM 生成答案

| 模組 | 技術 | 說明 |
|---|---|---|
| 文本切塊 | RecursiveCharacterTextSplitter | 滑動視窗 size=500, overlap=100 |
| Dense Search | Qdrant + Embedding（4096 維） | 語意向量搜尋 Top-10 |
| Sparse Search | BM25Okapi + jieba 中文分詞 | 關鍵字搜尋 Top-10 |
| Hybrid Fusion | Reciprocal Rank Fusion | k=60，融合兩種搜尋結果 |
| ReRank | LLM 相關性評分（0-10 分） | 重新排序取 Top-3 |
| Query ReWrite | LLM 口語化 → 正式語句 | 處理代名詞、口語表達 |
| Answer Gen | gemma-3-27b-it | 從 Top-3 段落萃取精準答案 |
| Evaluation | DeepEval（5 項指標） | 自動化品質評估 |

## 二、DeepEval 實際評估結果

從 30 題中隨機抽樣 5 題（Q1, Q4, Q9, Q21, Q24），使用 DeepEval 評估 5 項指標。評估模型為課程提供的 gemma-3-27b-it，透過自訂 DeepEvalBaseLLM 類別整合。

### 各題評估分數

| 題號 | Faithfulness | Answer Rel. | Ctx Recall | Ctx Precision | Ctx Relevancy |
|---|---|---|---|---|---|
| Q1 | 1.0000 | 0.8235 | 1.0000 | 0.8333 | 0.6471 |
| Q4 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.8462 |
| Q9 | 1.0000 | 1.0000 | 1.0000 | 0.9500 | 0.9167 |
| Q21 | 1.0000 | 1.0000 | 1.0000 | 0.7500 | 0.6923 |
| Q24 | 1.0000 | 0.8462 | 1.0000 | 1.0000 | 0.2222 |

平均分數總覽

| 指標 | 平均分數 | 評等 | 說明 |
|---|---|---|---|
| Faithfulness | 1.0000 | 完美 | 答案完全忠於 context，零幻覺 |
| Answer Relevancy | 0.9339 | 優秀 | 答案高度切中問題主題 |
| Contextual Recall | 1.0000 | 完美 | 所有關鍵資訊都被檢索到 |
| Contextual Precision | 0.9067 | 優秀 | 相關段落排序精準 |
| Contextual Relevancy | 0.6649 | 尚可 | 部分 context 包含雜訊段落 |

## 三、DeepEval 五項評估指標說明

### （1）Faithfulness 忠實度 — 平均 1.0000

定義：衡量 LLM 生成的答案是否「忠於」所檢索的上下文資料。若答案中包含未出現在 context 中的資訊（幻覺），則分數下降。

計算方式：DeepEval 將答案拆解為多個「claims」（事實宣稱），逐一檢查每個 claim 是否能從 retrieval_context 中找到支持。分數 = 有支持的 claims / 總 claims。

本系統結果：5 題全部拿到滿分 1.0，代表 LLM 完全沒有產生幻覺。歸功於 System prompt 強調「只根據參考資料回答」以及低 temperature=0.1。

### （2）Answer Relevancy 答案相關性 — 平均 0.9339

定義：衡量 LLM 的答案是否與用戶的問題相關。即使答案內容正確，若偏離問題主題，分數也會低。

計算方式：DeepEval 從答案中生成多個假設性問題（hypothetical questions），計算這些假設問題與原始問題的語意相似度。

本系統結果：Q4、Q9、Q21 拿滿分，Q1（0.82）和 Q24（0.85）略低，可能因為回答包含了額外的補充資訊。整體平均 0.93 表現優秀。

### （3）Contextual Recall 上下文召回率 — 平均 1.0000

定義：衡量檢索到的 context 是否包含回答問題所需的「所有」關鍵資訊。需要「參考答案」（expected_output）來計算。

計算方式：DeepEval 將參考答案拆解為多個 claims，檢查每個 claim 是否出現在 retrieval_context 中。

本系統結果：5 題全部滿分 1.0，代表 Hybrid Search（Dense + BM25）+ RRF 融合成功檢索到所有關鍵資訊。BM25 補強了精確關鍵字匹配，Dense 補強了語意理解。

### （4）Contextual Precision 上下文精確度 — 平均 0.9067

定義：衡量「相關的」context 是否排在「不相關的」前面。即使檢索到了正確資訊，若排序不佳，分數也會低。

本系統結果：Q4 和 Q24 拿滿分，Q21（0.75）相對較低。ReRank 機制有效提升了相關段落的排序，整體平均 0.91 表現優秀。

### （5）Contextual Relevancy 上下文相關性 — 平均 0.6649

定義：衡量檢索到的 context 整體上與問題的相關程度。若 context 中有大量與問題無關的雜訊段落，分數會低。

本系統結果：Q9（0.92）和 Q4（0.85）表現好，但 Q24（0.22）明顯偏低。Q24 的問題是「不喜歡紙張單子，有沒有別的方式」，口語化程度高，即使 Query ReWrite 處理後仍檢索到較多雜訊段落。此為主要改善方向。

## 四、基於 DeepEval 結果的 RAG 優化策略

| DeepEval 指標 | 本系統分數 | 對應的優化策略 |
|---|---|---|
| Faithfulness (1.0000) | 滿分 | System prompt 限制「只根據參考資料」+ temperature=0.1 |
| Answer Relevancy (0.9339) | 優秀 | Query ReWrite 口語→正式 + prompt 強調「直接回答」 |
| Contextual Recall (1.0000) | 滿分 | Hybrid Search (Dense+BM25) 互補 + chunk_size=500 |
| Contextual Precision (0.9067) | 優秀 | ReRank (LLM 0-10分) + RRF k=60 融合排序 |
| Contextual Relevancy (0.6649) | 待優化 | ReRank 篩選 Top-3 + ReWrite 提升查詢精確度 |

### （1）Query ReWrite 優化

用戶提問多為口語化（如「白白的像牛奶」、「那個紙張的單子」），直接搜尋會導致向量匹配不佳。透過 LLM ReWrite 轉為正式術語（如「自來水白濁 空氣混入 氣泡」、「紙本帳單電子帳單」），大幅提升 Dense Search 的召回率，間接提升 Contextual Recall 和 Precision。

### （2）Hybrid Search 優化

Dense Search 擅長理解語意（「水費很奇怪」→ 找到「水費異常」相關段落），BM25 擅長精確匹配關鍵字（「OTP」、「1910」等專有名詞）。兩者透過 RRF 融合互補，使 Contextual Recall 達到滿分 1.0。

### （3）ReRank 優化

Hybrid Search 返回 Top-10 候選段落後，使用 LLM 逐一評估相關性（0-10 分），重新排序後只取 Top-3。同時提升了 Contextual Precision（0.91）和 Contextual Relevancy（過濾雜訊）。

### （4）Contextual Relevancy 改善方向

此指標平均 0.6649 是最低的一項。分析 Q24 發現，高度口語化的問題（「那個紙張的單子」）即使經過 ReWrite，仍會檢索到部分不相關的段落。未來改善方向包括：（a）更積極的 ReRank 門檻過濾（如 score < 5 直接丟棄）；（b）減少 Top-K 從 3 到 2；（c）多輪 ReWrite 提升查詢品質。

## 五、DeepEval 技術整合方式

DeepEval 預設使用 OpenAI GPT 系列模型，本系統透過繼承 DeepEvalBaseLLM 類別，自訂使用課程 LLM API（gemma-3-27b-it），並加入重試機制處理 API Timeout：

### 核心實作要點

（a）CustomLLM 類別：繼承 DeepEvalBaseLLM，覆寫 generate() 方法，內部透過 OpenAI SDK 連接 ws-02.wade0426.me/v1 API。

（b）指數退避重試：遇到 524 Timeout 時，等待 10→20→40→80→160 秒自動重試，最多重試 5 次。同時偵測 HTML 錯誤頁面回傳。

（c）斷點續跑：RAG 結果和 DeepEval 評估分數分別儲存為 JSON checkpoint，中途 crash 後重新執行可自動從斷點繼續，不需重跑已完成的部分。

### 評估所需的輸入資料

| DeepEval 參數 | 對應資料 | 說明 |
|---|---|---|
| input | 用戶問題 | 30 題口語化問題（抽樣 5 題） |
| actual_output | RAG 生成的答案 | 系統回答 |
| expected_output | 參考標準答案 | questions_answer.csv |
| retrieval_context | Top-3 檢索段落 | ReRank 後的段落列表 |

## 六、總結

本系統透過 DeepEval 的五項指標量化評估 RAG 品質，實際結果顯示：

（1）Faithfulness = 1.0 和 Contextual Recall = 1.0 達到滿分，證明系統不會產生幻覺，且檢索能力完整。

（2）Answer Relevancy = 0.93 和 Contextual Precision = 0.91 均達優秀水準，Query ReWrite 和 ReRank 機制發揮了重要作用。

（3）Contextual Relevancy = 0.66 為最低指標，反映高度口語化問題仍會引入部分不相關的 context，是未來優化的主要方向。

整體而言，DeepEval 提供了客觀、可量化的評估框架，使我們能夠精準定位 RAG 系統各環節的瓶頸。相比人工評估，DeepEval 更具效率和一致性，是建構生產級 RAG 系統不可或缺的工具。