我使用了三個函數分別是 calculate_faithfulness、calculate_answer_relevancy 及 calculate_contextual_metrics，模擬了 DeepEval 框架的評分邏輯，透過 LLM 作為裁判，針對生成答案的忠實度與相關性進行評分。

calculate_faithfulness:是用來檢測 AI 是否在胡說八道的

calculate_answer_relevancy: 是用來檢查 AI 是否有準確回答使用者的問題

calculate_contextual_metrics: 這是用來測量搜尋到的資料到底好不好用

根據生成的 csv 檔顯示，五題當中第一題的 Faithfulness 僅僅拿到了 0.6 分，因此加了 Rerank 來讓 AI 檢索上下文且調整生成端的限制，後續 2~5 題的分數都在 0.95 以上以完成優化。