評分方式

本次複賽將藉由「GPT-4 自動化評分」為依據，去進行 RAG 生成結果的評分，公式如下：

$$\text{Automatic Evaluation Score} = \frac{2\times \text{n\_correct} + \text{n\_miss}}{\text{n}} - 1$$

其中

- n: 總題數
- n_correct: 被 GPT-4 判讀為正確預測的題數
- n_miss: 該題以「不知道」作為應答的題數 (字串需完全相符)

同時從公式中，亦可推演得到

- 不正確預測: 即 n - n_correct，被 GPT-4 判讀為不正確預測的題數

複賽將以 **Automatic Evaluation Score**（將四捨五入計算至小數點後 7 位）作為成績，範圍為 -1 ~ 1 之間，詳細可以參考以下範例。

範例一

問題	生成結果	真實標記	GPT-4 判讀
3 分鐘 05 秒是多少秒？	3 分鐘 05 秒是 185 秒。	185 秒	True
請問玉山銀行的總部位於台灣的哪個縣市？	玉山銀行的總部位於新竹市。	台北市	False
請問在2023年第三季的聯電財務報告中，哪些公司被列為聯電的關聯企業及其他關係人？	不知道	智源科技, 新興電子	(False)
一年四季是指哪四季？	四季是指春、夏、秋、冬	春夏秋冬	True

- n: 4
- n_correct: 2
- n_miss: 1
- Automatic Evaluation Score: $\frac{2\times 2+1}{4} - 1 = 0.25$

範例二

問題	生成結果	真實標記	GPT-4 判讀
一週是幾天？	365天	7 天	False
請問台灣最高的山是哪座？	玉山	玉山	True
美國的首都位於哪裡？	我不知道	華盛頓哥倫比亞特區	False
棒球比賽每一局有幾人出局？	不知道	三人出局	(False)

- n: 4
- n_correct: 1
- n_miss: 1
- Automatic Evaluation Score: $\frac{2\times1+1}{4}-1=-0.25$

注意事項

- 本次評分程式碼由於採機器自動化評估，因此主辦提供評分程式碼讓參賽者可以於線下環境、不受限之雲端環境，進行 RAG 生成評測的準則的依循。
- 此開源評分程式碼旨在提供參賽者有生成結果的評分依循的方向，但由於提示 prompt 或外部 API 資源調度等因素，此程式碼產生的評分結果，可能與主辦單位於正式複賽使用的評分程式碼評分結果，存在微幅差異。
- 主辦單位不會提供評分程式碼所使用之 OpenAI GPT-4 API 資源。
- 參賽者請確保繳交結果格式參照本次複賽的資料集說明文件，在此列出可能的作答狀況
    - 若繳交之「整檔」格式錯誤，導致評估程式碼無法解析該檔案 (例如: JSONDecodeError)，此次 Automatic Evaluation Score 將直接以 -1 作為最終成績。
    - 若繳交之「該題」超出本次複賽生成之 400 token limit，該題將以上述提及「不正確預測」作為計數累加。
    - 若繳交之「該題」格式錯誤，導致該題無法判定 (例如: KeyError)，該題將以上述提及「不正確預測」作為計數累加。