複賽資料說明

本次複賽使用網路上供社會大眾查閱的公開資料，如下說明

資料類別 (`category`)	簡介	參考文件 (`reference`) 檔案類型
faq	玉山銀行官方網站上的常見問題	.json
insurance	玉山銀行代銷的保險產品之保單條款	.pdf
finance	公開資訊觀測站上的上市公司財務報告	.pdf

下圖表示了複賽的資料夾結構:

```
├── dataset
│   ├── final
│   │   └── questions_example.json
│   │   └── ground_truths_example.json
└── reference
    ├── faq
    │   └── pid_map_content.json
    ├── insurance
    │   ├── 1.pdf
    │   ├── 2.pdf
    │   └── ...
    └── finance
        ├── 0.pdf
        ├── 1.pdf
        └── ...
```

- `questions_example.json` 為範例題目，共有 150 題，供參賽者練習使用。
- `ground_truths_example.json`: 範例題目的答案，供參賽者練習使用。
- 資料夾 `reference` 存放各類型資料的參考文件

資料格式-複賽

1. 主辦單位發佈的題目格式 (`questions_example.json`)

`questions_example.json` 的格式與正式複賽題目 `questions_final.json` 及串測題目 `questions_test.json` 相同

```
{
    "questions": [
        {
            "qid": 1,
            "source": [442, 115, 440, 196, 431, 392, 14, 51],
            "query": "匯款銀行及中間行所收取之相關費用由誰負擔?",
            "category": "insurance"
        },
        // 後面題目省略...
    ]
}
```

題目包在 `questions` 串列裡，每個題目皆有:

欄位	型態	說明
qid	integer	題號
query	string	問題
source	list of integer	能夠回答問題的可能選項，數字的意義為文件編號 (`pid`)，可在資料夾 `reference` 中找到對應的檔案或內容，例如: - 當 category 為 `insurance` 或 `finance`: pid 13 對應到文件庫中 `13.pdf` - 當 category 為 `faq`: pid 13 對應到 `pid_map_content.json` 中, 鍵為 13 的內容
category	string	資料類型，`reference` 裡有對應的資料夾存放該類型的文件

各 category 中，題目選項數量 (亦即 source 長度) 最大值為:

- insurance: 9
- finance: 9
- faq: 16

2. 參賽者繳交的答案格式 (複賽請命名為 `final_pred.json`，串測請命名為 `test_pred.json`)

```
{
    "answers": [
        {
            "qid": 1,
            "generate": "由匯款人負擔"
        },
        // 後面題目省略...
    ]
}
```

參賽者繳交的預測結果放在 `answers` 串列裡，每個題目皆有:

欄位	型態	說明
qid	integer	題號
generate	string	回答問題的答案文字

繳交的答案文字長度限制在 400 字符 (token) 以下，計算方式可參考這裡，例如:

```
encoding = tiktoken.encoding_for_model("gpt-4")
```