

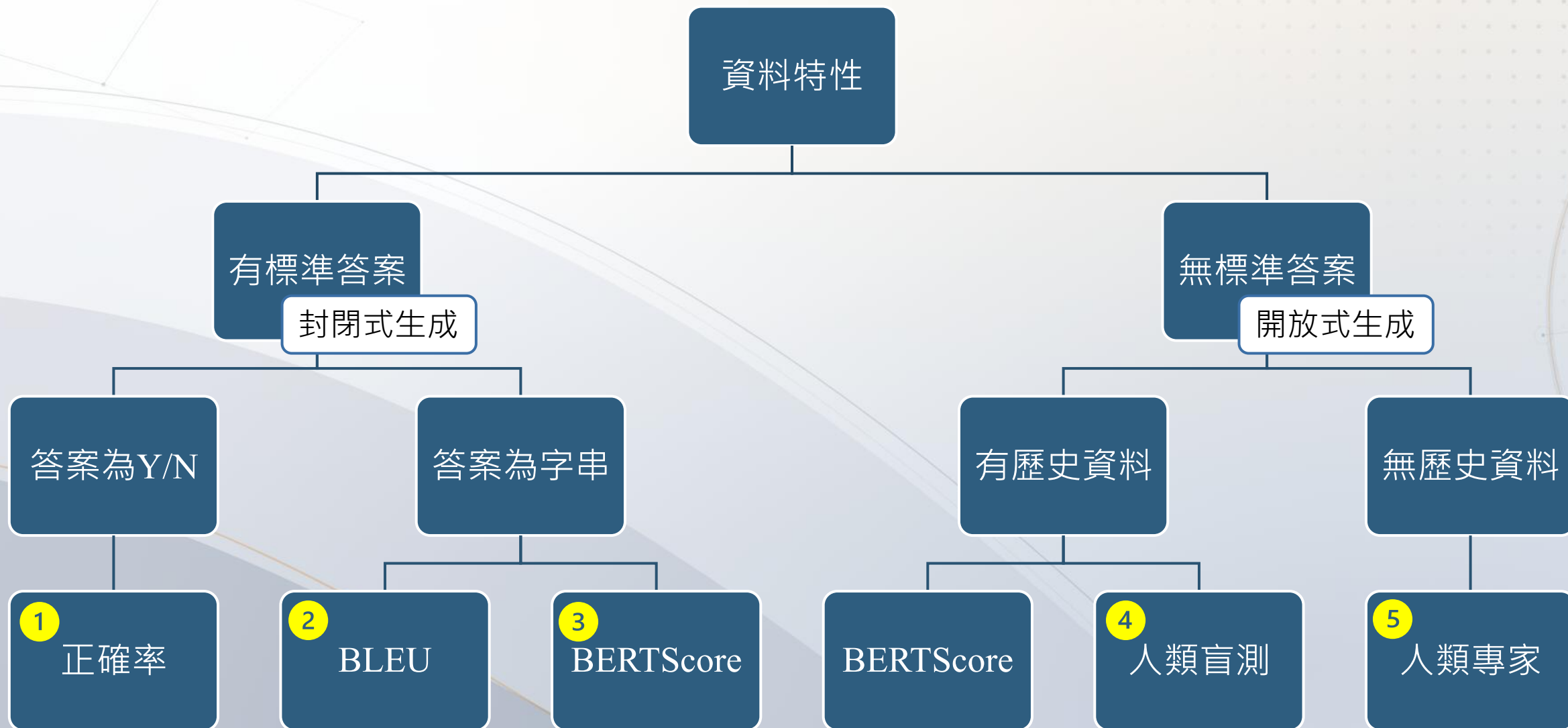
評 - 依據資料特性規劃評量方式

* 目前在規劃階段，只有進行訓練集的評量

115年7月14日

特別感謝 佳蕙、晉傑、人豪

結論：依資料特性選用不同測評



提案3：職災報告書實例智能輔助撰寫

熊圖建設股份有限公司「敘蒔美大樓新建工程」之再承攬人潘○○所僱印尼籍移工HAMZANI發生墜落災害致死重大職業災害檢查報告書

一、事業單位概況：（發生職業災害之事業單位）

（一）行業分類：其他專門營造業（4390）

（二）事業單位名稱：潘○○

（三）事業單位地址、電話：○○○、○○○

（四）營業人統一編號：無

（五）雇主：

1、事業主：

名稱：潘○○

地址：○○○

代表人職稱：負責人

姓名：潘○○

住址：○○○

2、事業經營負責人：

（1）經營負責人：同事業主代表人

（2）代理人：無

（六）工作場所負責人：同事業主代表人

（七）現場作業主管

1、部門主管：同事業主代表人

2、作業主管：未設置模板支撐作業主管

（八）勞工人數：男5人，女1人

（九）參加勞工職業災害保險人數：男0人，女0人

（十）參加勞工保險人數：男0人，女0人

行紀錄或

號2樓、

標題

勞工從事大樓新建工程作業發生墜落災害致死重大職業災害

罹災情形

死亡1人

發生經過

（一）民國115年2月2日，某○公司。

（二）7時30分許，潘**所僱勞工張**與張**帶著4名印尼籍移工至工地現場從事模板物料傳遞作業，將5樓模板物料傳遞至頂樓。張**與張**及1名移工於頂樓作業，餘3名移工在5樓作業，其中2名移工於5樓室內整理物料並傳遞給5樓施工架上的移工H**，H**將物料向上傳遞給在頂樓施工架的另1名移工，再交由張**與張**整理。作業至11時許，作業位置移到5樓C戶，張**突然聽到移工呼喊，才發現H**墜落至1樓，張**隨即趕至1樓將其送往○○○醫院治療，隨後轉往○○○醫院住院治療，仍於同月5日13時13分不治死亡。

正確率 以應用功能3計算罹災情形為例

完全匹配率 EM

- AI 算出的「死亡人數」與「受傷人數」跟標準答案完全一模一樣的報告篇數，除以總篇數

$$EM = \frac{AI \text{ 完全算對的篇數}}{\text{總測試篇數}} \times 100\%$$

平均絕對誤差 MAE

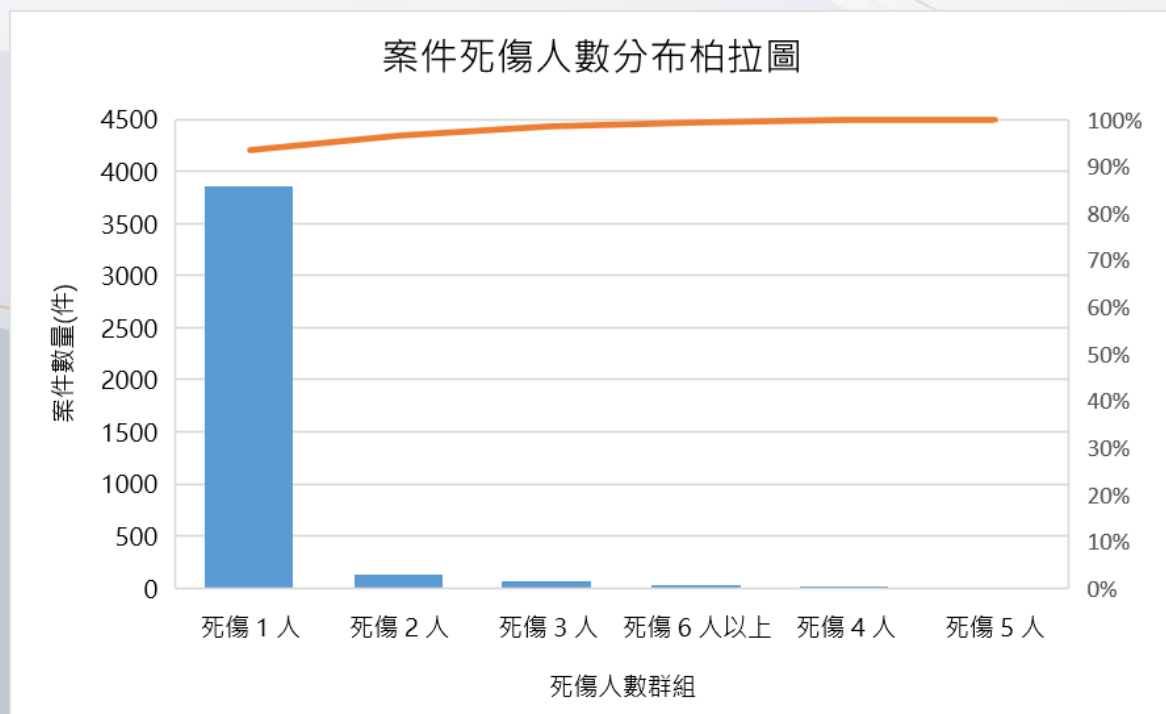
- 計算 AI 算出的數字與真實數字的「差距絕對值」並取平均值

$$MAE = \frac{\text{所有報告誤差人數總和}}{\text{總測試篇數}}$$

正確率

以應用功能3計算罹災情形為例

	真實有死傷人數 (正樣本)	真實無死傷/人數為0 (負樣本)
AI 算出的數字正確	真正例 (True Positive, TP) 有死傷且 AI 算對了。	真負例 (True Negative, TN) 沒死傷且 AI 也算對了。
AI 算出的數字錯誤	偽負例 (False Negative, FN) 實際有死傷，但 AI 算錯 (或漏抓) 。	偽正例 (False Positive, FP) 實際沒死傷，但 AI 瞎掰了一個數字。



本例屬全正樣本
故不適合使用混淆矩陣

→死傷1人達93%，不考慮誠信問題的話
好像我錯誤訊息直接寫死亡1人也很容易正確 😂

BLEU 以應用功能4提取相驗屍體證明書為例

相驗屍體證明書

地檢署撰寫

LLM地端模型

Gemma 4

AI 提取

實際 AI 回答

人類提取

預期 AI 回答

兩者比較

BLEU 雙語評估研究指標 / 字面重合度

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002).

BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).

BLUE(%)	描述
60 – 100	關鍵名詞字面與標準答案幾乎完全一致
40 – 59	死因與醫學關鍵字高度匹配，無漏字偏誤
30 – 39	核心詞彙正確匹配，但有輕微格式或贅字差異
10 – 29	錯失關鍵死因名詞，或擷取到無關干擾文字
低於 10	無任何重合字詞，無法用於正式相驗報告

**BLEU 分數達 40 以上
視為 AI 可精確萃取法律與醫學關鍵專名**

BLEU 以應用功能4提取相驗屍體證明書為例

BLEU追求字字對齊，建議搭配BERTScore使用

期待AI回答	實際AI回答	BERTScore_F1	BLEU_100
車禍(機車-自小客車)	車禍-機車自小客車	0.8956	33.2587
中樞神經休克	中樞神經炎	0.8689	54.7518
顱骨骨折併顱內出血及肋骨骨折	顱骨骨折內出血及肋骨骨折	0.9413	70.6349
高處墜落	高嚴重跌落	0.7284	12.7033
全身多處骨折及胸部骨折壓迫窒息	全身多處骨折及胸部骨折迫窒息	0.9673	80.5225
○○鋸	○○鉉	0.8884	55.0321

BLEU 以應用功能4提取相驗屍體證明書為例

基於 **N-gram (連續字詞組合)** 的重合度
但分母定義完全不同

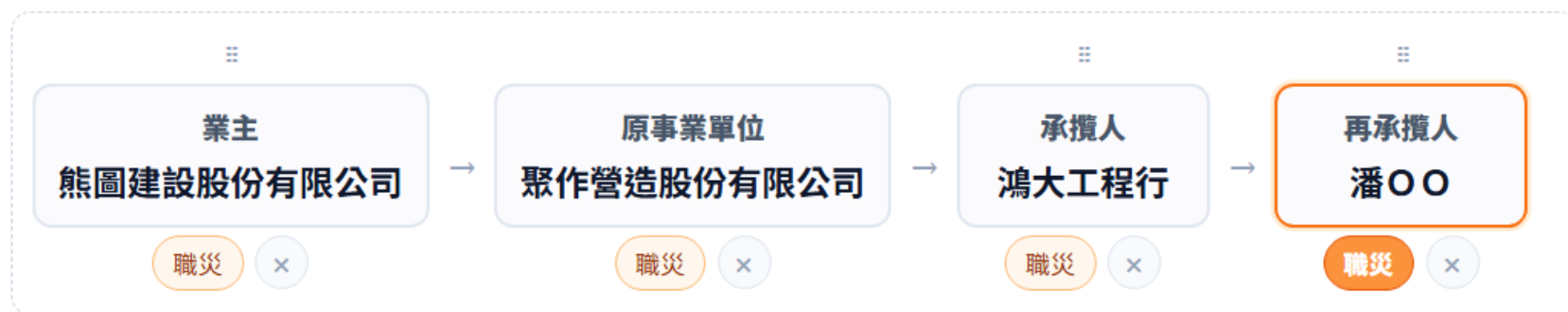
評估情境	建議選用指標	方法學理由
絕對不容許漏抓關鍵 例如：漏抓致命傷、漏抓併發症	ROUGE	Recall 導向 標準答案中的關鍵醫學名詞 百分之百都有被 AI 撈出來
絕對不容許 AI 瞎編/幻覺 例如：沒發生的死因自己腦補	BLEU	Precision 導向 確保 AI 寫出來的字字都有依據 不能夾帶幻覺廢話



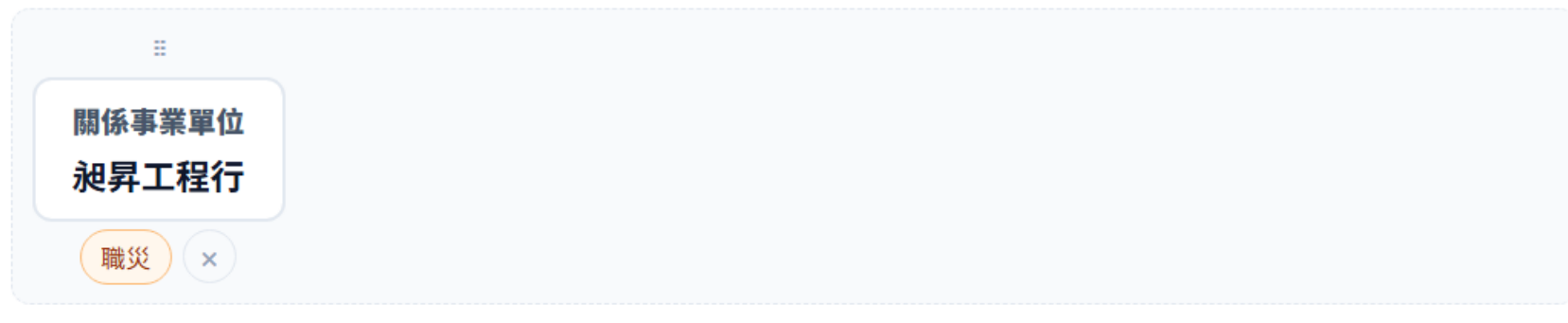
提案4：職災報告書智能輔助撰寫

將已完成之實體關係
撰寫為報告書標題

承攬關係鏈



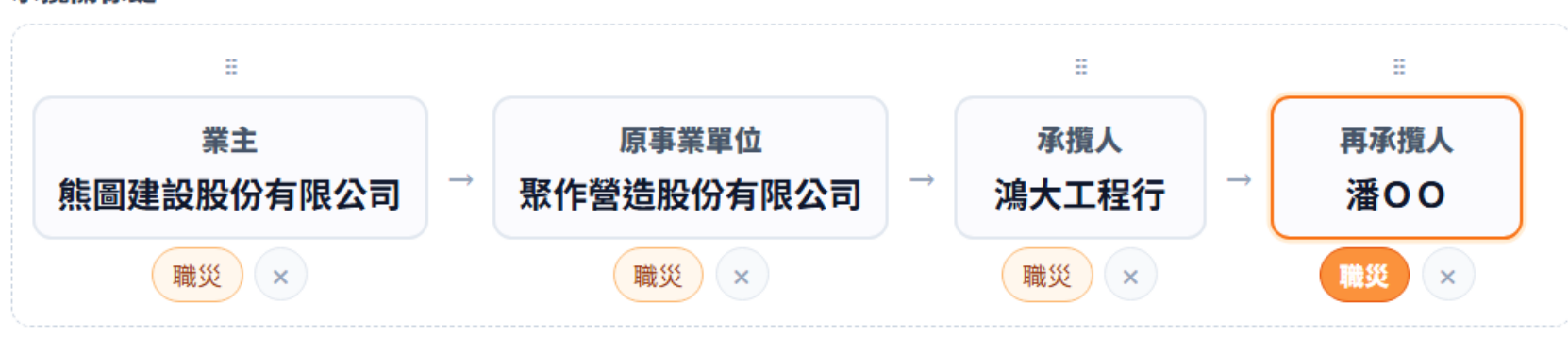
關係事業



熊圖建設股份有限公司「敘蒔美大樓新建工程」之再承攬人潘〇〇所僱勞工H〇〇發生墜落、滾落災害致死重大職業災害檢查報告書

提案4：職災報告書智能輔助撰寫

承攬關係鏈



**將已完成之實體關係
撰寫為承攬關係**

1. 熊圖建設股份有限公司將「敘蒔美大樓新建工程」（以下簡稱本工程）全部交付聚作營造股份有限公司承攬，以總價新臺幣〔工程總價〕元整（〔含稅或未稅〕）〔發包條件〕，雙方訂有〔契約名稱〕，認定熊圖建設股份有限公司為業主。
2. 聚作營造股份有限公司將本工程之「〔工程細項或施作項目〕」，以總價新臺幣〔工程總價〕元整（〔含稅或未稅或無〕）交由鴻大工程行承攬，雙方訂定有〔契約名稱〕。
3. 鴻大工程行將本工程之「〔工程細項或施作項目〕」，以總價新臺幣〔工程總價〕元整（〔含稅或未稅或無〕）交由潘〇〇承攬，雙方訂定有〔契約名稱〕。
4. 本案罹災者H****為潘〇〇所僱之勞工。
5. 綜上，本案以熊圖建設股份有限公司為業主，聚作營造股份有限公司為原事業單位，以鴻大工程行為承攬人，以潘〇〇為再承攬人，以昶昇工程行為關係事業單位，撰寫本重大職業災害檢查報告書。

BERTScore 以應用功能3撰寫職災實例發生經過為例

重大職業災害檢
查報告書

人類勞檢員撰寫

LLM地端模型

Gemma 4

重大職業災害實
例

AI 撰寫

重大職業災害實
例

人類勞檢員撰寫

兩者比較

BERTScore 語意相關性評分

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi (2020). BERTScore: Evaluating Text Generation with BERT. International Conference on Learning Representations (ICLR).

F1	描述
0.90 – 1.00	幾乎完全語意一致
0.80 – 0.89	高度語意相似
0.70 – 0.79	中等語意相似，有部分語意差異
0.60 – 0.69	語意有限相似，理解上可能出現偏差
低於 0.60	語意不相似，可能無法互換或傳遞相同意涵

**F1 分數達 0.80 以上
視為 AI 可撰寫與人類高度語意相似內容**

BERTScore 以應用功能3撰寫職災實例發生經過為例

NO	人類勞檢員撰寫	AI 撰寫	F1
1	(一)民國115年1月8日，花蓮縣，中○公司。 (二)罹災者於石灰石堆置區從事原物料搬運工程之重機械鏟挖作業，罹災者為挖掘機司機於該區操作挖掘機將石灰石搬運至聯結車上方，於作業過程中，罹災者離開挖掘機駕駛室，走入挖掘之料堆防護溝(高風險區域)，此時上方石灰石發生崩塌滑落，致罹災者遭掩埋。	(一)民國115年1月8日12時50分許，花蓮縣，亞泥花蓮廠，中○公司。 (二)中○公司所僱勞工潘**於亞泥花蓮廠內之石灰石堆置區從事原物料搬運工程之重機械鏟挖作業。勞工潘**及另一名勞工陸**均為挖掘機司機，當時各自操作一輛挖掘機將石灰石搬運至聯結車上方。於作業過程中，潘**離開挖掘機駕駛室，走入挖掘之料堆防護溝（高風險區域），此時上方石灰石發生崩塌滑落，致潘**遭掩埋。於13時18分許，陸**接獲來裝車的聯結車司機告知另一挖掘機無人，故前往尋找潘**，發現現場有崩塌痕跡，便操作挖掘機及徒手尋找；當尋獲潘**遺體後，以徒手將其拉出石灰石堆，惟已死亡。	0.8026

人類單盲測試

以應用功能3撰寫職災實例發生經過為例

1. 前置作業

- 去識別化
- 混編分配

2. 專家審查

- 至少 3位
不同轄區
之勞檢員
專家

3. 測評標準

- 無須修改
- 小幅修改
- 大幅修改

(後續將提供詳盡評分依據)

4. 解盲驗證

- 替代採用
率達 85%
以上

$$\text{替代採用率} = \frac{\text{AI組 (無須修改+小幅修改) 篇數}}{\text{AI組總測試篇數}} \times 100\%$$

人類單盲測試

以應用功能3撰寫職災實例發生經過為例

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
北區中心	AI	AI	AI	人類	人類	人類
	人類	人類	人類	AI	AI	AI
南區中心	AI	AI	AI	人類	人類	人類
	人類	人類	人類	AI	AI	AI

1. 避免淪為挑剔→每1案例只會拿到1份不知道是誰寫的文本
2. 減輕人類負擔→每1文本只要回答1個問題（無須 / 小幅 / 大幅修改）
3. 避免轄區差異→每一轄區只會拿到隸屬於該轄區的案件
4. 避免樣本不夠→第1次沒過的話我再拿第1次沒看過的樣本測第2次
（如果真的不夠資料做第2批測試集才這樣做）
5. 避免過於嚴苛→如果AI測評不到85%，拿勞檢員對於人類的回饋做比對

提案4：職災報告書智能輔助撰寫

五、災害發生經過：

依據中國鋼鐵結構股份有限公司工地主任陳**、現場勞工A塔塔吊(肇災倒塌之塔吊)吊掛指揮手姚**、現場勞工B塔塔吊(鄰近受波及之塔吊)吊掛指揮手陳**及鎮鴻企業有限公司工地負責人賴**所述，113年11月18日在完成該塔吊爬升作業(座8層/夾11層)，113年11月22日再進行鋼梁拆除作業，當日上午約9時25分，A塔塔吊操作手張**準備第3次吊掛拆除原座5層之鋼梁，於空載無荷重下將塔吊伸臂上升至最高仰角時，A塔塔吊就發生倒塌事件，罹災者張**被拋出駕駛室掉落至4樓樓板，經送醫不治於同日死亡。A塔塔吊倒塌時塔吊伸臂撞擊鄰近一座B塔塔吊，造成B塔塔吊操作手林**受傷送醫於同日返家休養，以及B塔14樓從事安全網清潔作業之越南籍勞工阮**受傷，經送衛生福利部雙和醫院住院治療。

AI偵測人機料法環，提供撰寫引導

 系統偵測到以下關鍵實體，請於【六、災害現場概況】補充對應之科學量測鐵證：

【人】人員狀態：

- 罹災者張**被拋出駕駛室掉落至4樓樓板，請提供罹災者遺體發現之確切座標位置、墜落高度（從駕駛室地板至4樓樓板之垂直距離）及墜落點與塔吊基座之水平距離。

【機】機具設備：

- A塔塔吊之確切型號、出廠年份、最大起重量（SWL）及本次事故發生時之規格參數。

【料】物料狀況：

- 事故發生時，A塔塔吊雖為「空載」，但請確認其伸臂上是否有殘留之吊具（如：吊鉤、鋼索、吊帶）及其材質與規格。

提案4：職災報告書智能輔助撰寫

五、災害發生經過：

依據中國鋼鐵結構股份有限公司工地主任陳**、現場勞工A塔塔吊(肇災倒塌之塔吊)吊掛指揮手姚**、現場勞工B塔塔吊(鄰近受波及之塔吊)吊掛指揮手陳**及鎮鴻企業有限公司工地負責人賴**所述，113年11月18日在完成該塔吊爬升作業(座8層/夾11層)，113年11月22日再進行鋼梁拆除作業，當日上午約9時25分，A塔塔吊操作手張**準備第3次吊掛拆除原座5層之鋼梁，於空載無荷重下將塔吊伸臂上升至最高仰角時，A塔塔吊就發生倒塌事件，罹災者張**被拋出駕駛室掉落至4樓樓板，經送醫不治於同日死亡。A塔塔吊倒塌時塔吊伸臂撞擊鄰近一座B塔塔吊，造成B塔塔吊操作手林**受傷送醫於同日返家休養，以及B塔14樓從事安全網清潔作業之越南籍勞工阮**受傷，經送衛生福利部雙和醫院住院治療。

AI偵測人機料法環，提供撰寫引導

 系統偵測到以下關鍵實體，請於【六、災害現場概況】補充對應之科學量測鐵證：



【法】作業動態：

- 發生倒塌瞬間，A塔塔吊伸臂之「確切仰角」度數。
- 塔吊爬升作業（座8層/夾11層）完成後，是否有進行結構穩定性檢查或紀錄？
- 事故發生前，該塔吊之定期檢查紀錄、維修紀錄及操作手之合格證書紀錄。



【環】環境變異：

- A塔倒塌後，其結構組件（如塔身、伸臂）撞擊地面或鄰近B塔之範圍與佔地面積。
- 請量測A塔與B塔之間之「絕對水平距離」，以判定受波及之物理力學關係。
- 現場地面或4樓樓板之受撞擊損壞狀況（如：裂縫、凹陷深度）。

人類專家評分

以應用功能4引導撰寫罹災現場概況為例

專家人選

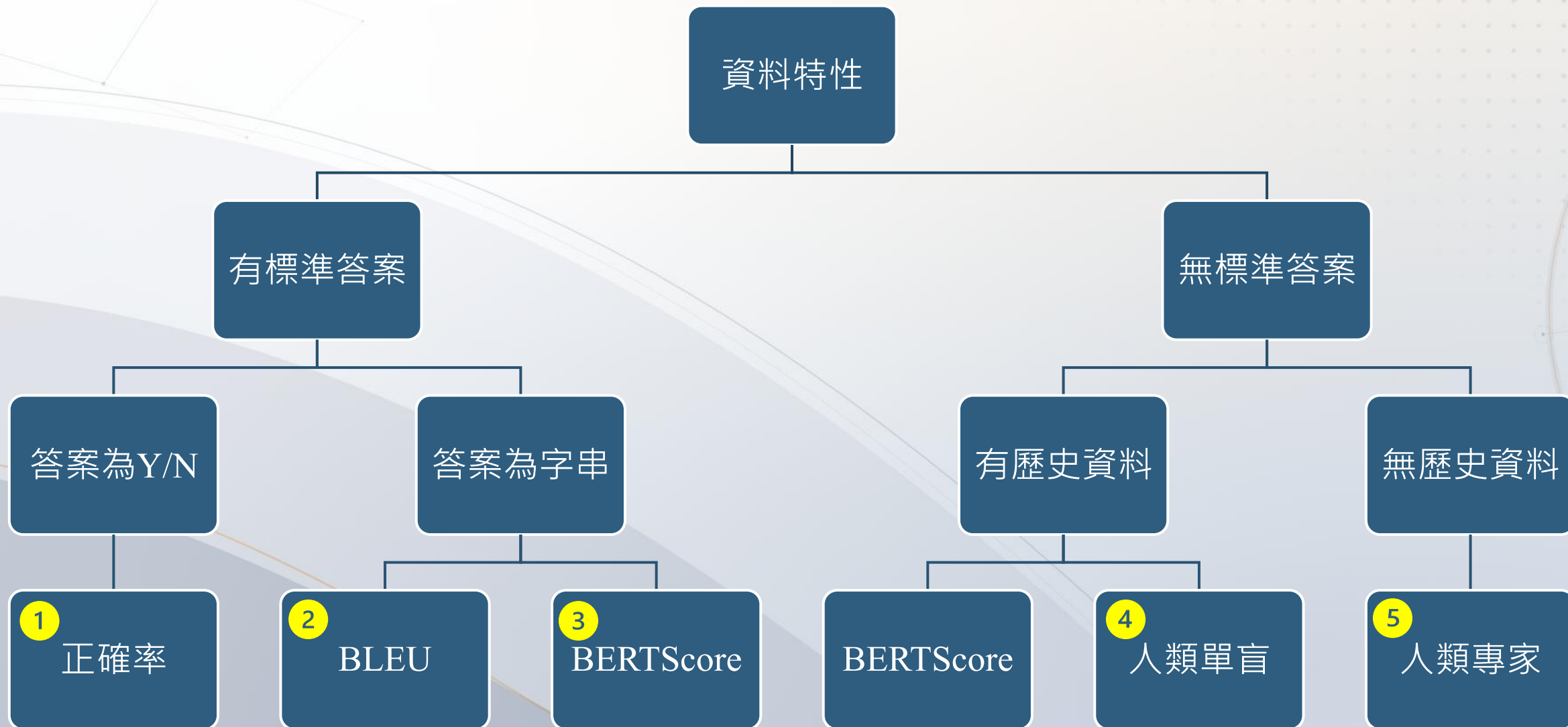
- 蘇國棟博士
- 現任國立宜蘭大學環境工程學系助理教授
- 曾任職於職安署北區職安中心
- 具備20年勞檢經驗

評分標準設計

- AI 產出實用度評分：以量化分數進行評估
- 是否有明顯錯誤：標註包含AI 明顯錯誤內容，進行後續補正
- 專家簡評：以資深勞檢員眼光，補正 AI 未及之處

我不具備有領域知識，沒有辦法驗證AI生成內容
希望透過這次的人類專家評分，建立完整的 Rubric，未來就可以套用 LLM as a Judge

結論：依資料特性選用不同測評



AI專有名詞補充說明

全局最優 > 局部完美

過擬合 Overfitting

AI 透過「死記硬背」現有的課本答案，導致它雖然能在模擬考拿滿分，但只要題目換個字，它就會在真實戰場上完全崩潰。

泛化 Generalization

AI 具備「舉一反三」的能力，即使面對從未見過的全新未知狀況，依然能精確抓到核心語意並給出正確答案。

退化 Regression

AI 陷入「顧此失彼」的修補陷阱，為了解決少數語意模糊的「地獄怪題」，反而動搖核心邏輯，導致原本穩拿高分的基礎題集體錯誤。

工作日誌分享

三階段

訓練集

驗證集

測試集

- 1 先評一次 在訓練集、正式給業主之前，先自己做過一次確認可行，避免自己坑康自己，例如只做BLEU可能有點危險
- 2 自動化評 可自動化跑分者，最省力，優先使用；人類測評要考量是否有充足的回傳時間，要安排工期提前交付給人類專家
- 3 分階段評 同時有機器測評和人類測評的，可以訓練集、驗證集、測試集都做機器測評；人類測評則於測試集再做以減輕負擔與成本
- 4 預留資料 因為分成三集做，需要保留足夠的資料量，可以支撐至少2次的測試集比較保險