

地質災害 × 人工智慧

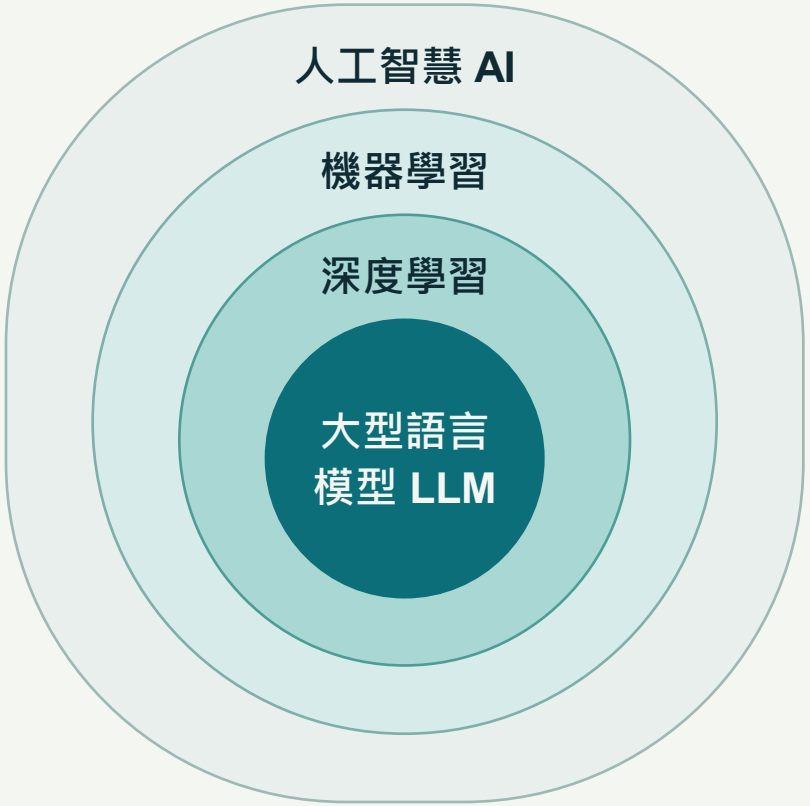
融合全球文獻與地端微調AI模型之堰塞湖知識庫檢索

張睿明｜AI X Earth 策展人｜2026.11.11

黃韋凱 中興工程顧問社
陳昆廷 屏東科技大學土木工程學系

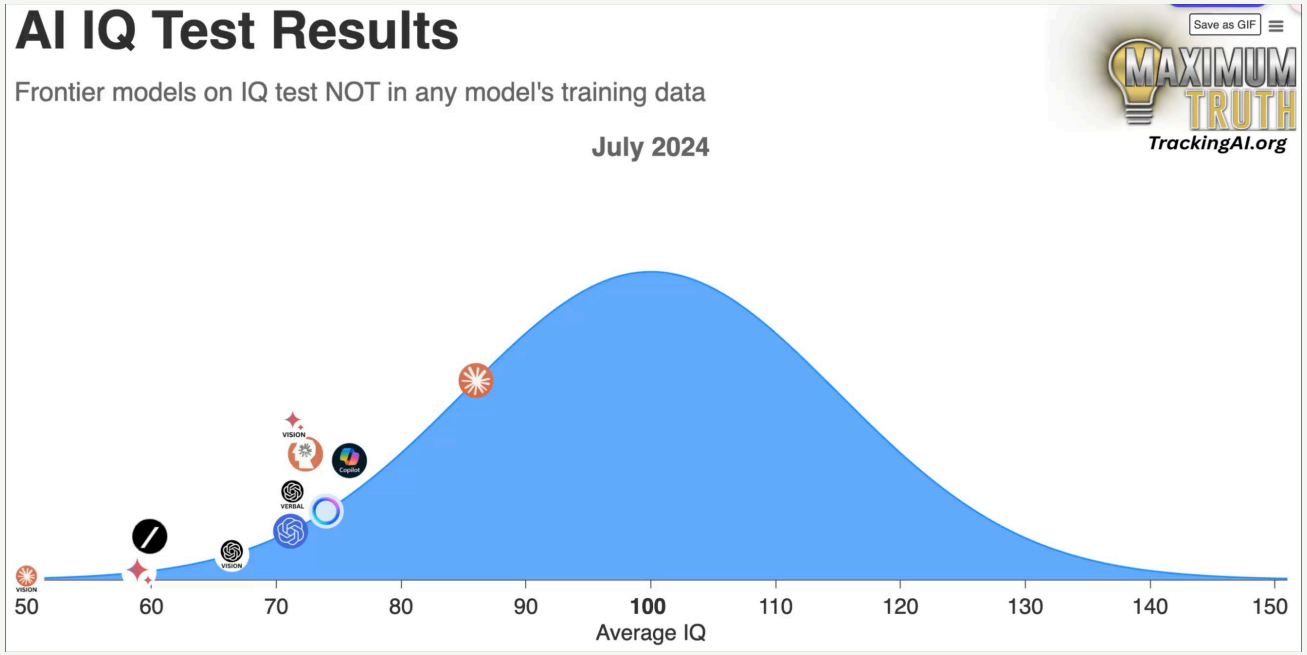
圖片來源:太管處

這裡說的 AI 是哪一層



今天談的是最內層：會讀、會寫大語言模型

像 ChatGPT、Claude 這類大型語言模型，幾年內從「偶爾能對話」進步到「能寫報告、寫程式」。





有些覺得很簡單的事，實際上非常困難。

叫 AI 去找正確的答案回來。

找得到．找得對．不能編．找不到要承認

RAG 的正式定義

RAG（Retrieval-Augmented Generation，檢索增強生成）：把模型權重裡的「參數式記憶」，與可替換、可更新的外部文件索引「非參數式記憶」結合；先檢索相關文件，再讓模型依這些文件產生答案。
Lewis et al., NeurIPS 2020

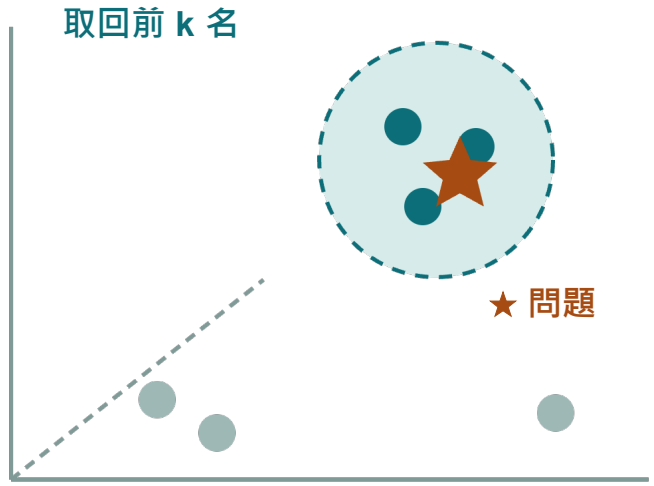
基礎 RAG 的三個步驟（Gao et al. 2023 綜述）

① 建索引
文件切成小塊，轉成向量存起來

② 檢索
問題也轉成向量，找最接近的 k 塊

③ 生成
問題 + 取回的小塊交給模型作答

模型看不到的內容，就答不出來：取回什麼，決定答案的上限。



示意圖；實際向量有 1,024 維（bge-m3 模型）

- 意思相近的句子：「壩高 130 公尺」「dam height of 130 m」「堰塞壩高度」

RAG 已用在各專業領域，但仍會出錯

醫療

Zakka et al., NEJM AI, 2024

評分：130 個臨床情境，5 位醫師評分

先查醫學資料再回答，正確性平均提升約 18%

法律

Magesh et al., J. Empirical Legal Studies, 2025

評分：研究者逐題判定回答是否含捏造內容

號稱「不會捏造」的商用工具，仍有超過 17% 的回答含捏造

地質災害

Wu et al., Frontiers in Earth Science, 2026

評分：60 題；8 位專家各自評分；每題 0–1 分（滿分 1 分）、評 6 個面向，取平均

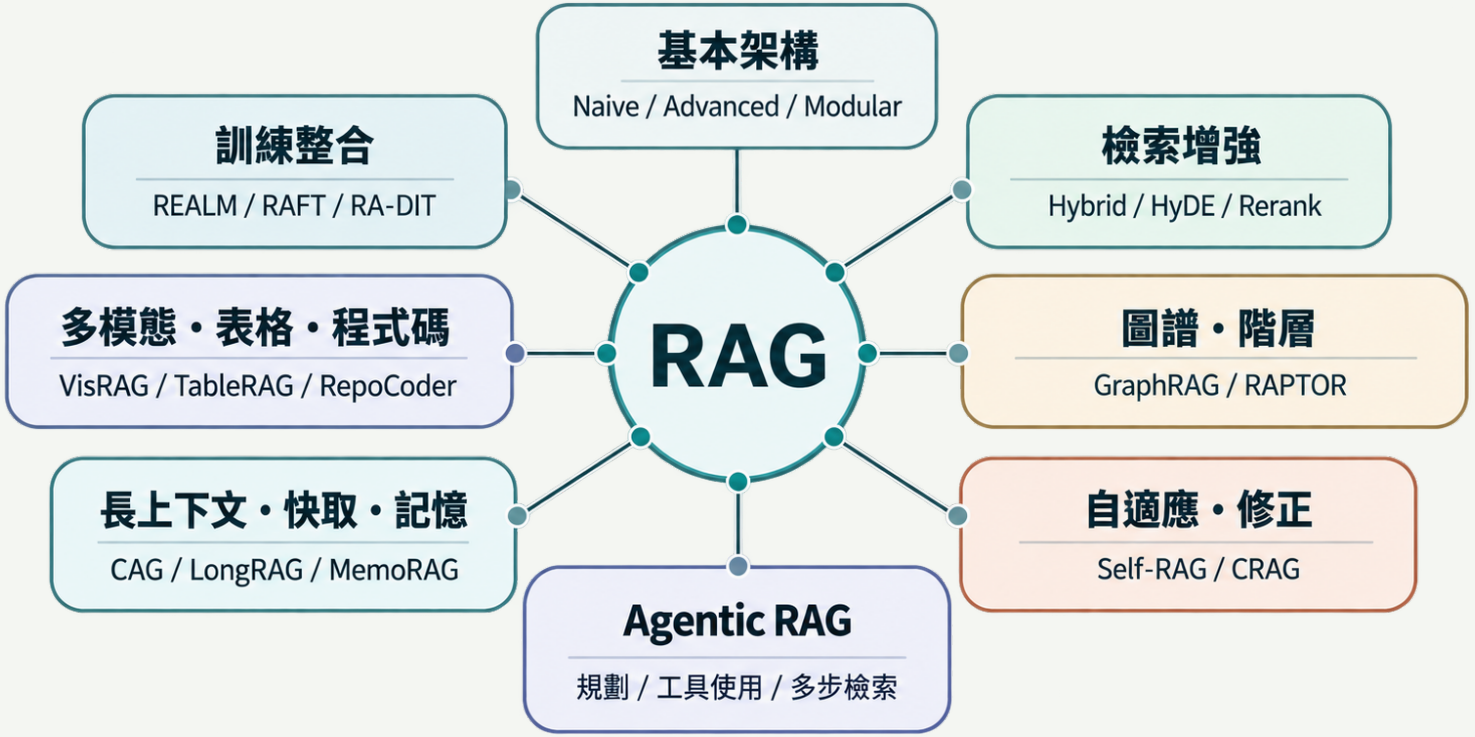
知識型題目 0.80–0.83；「批判思考」0.578、「創新延伸」0.550 最低

難在哪裡：資料整理與 RAG 的各種變形

資料整理

- PDF 轉成文字
- 表格圖中的文字
- 與公式
- 多語言與簡繁體
- 每一塊都要帶文獻與頁碼

這一步沒做好，後面什麼技術都救不回來。



沒有一種組合對所有資料都最好：只能用自己的資料、自己的題目反覆試，而且每次改動都要量化。

文獻，與今天要回答的三個問題

106
份國內外堰塞湖處置文獻

2,805
頁

426 萬
字



語言依各文獻正文的主要文字判定；不少繁中論文附有英文摘要與英文表格

問題 1 | 第 02、03 段

資料要怎麼整理、技術怎麼選，AI 才找得到？

問題 2 | 第 03、04 段

怎麼知道 AI 找對了、沒有亂編？

問題 3 | 第 04、05 段

一定要用最大的模型嗎？微調有沒有用？

四個步驟

輸入：106 份文獻

① 整理

切成 36,123 個帶頁碼的小塊；表格對照原圖校正

程式 + 校正

② 搜尋

關鍵詞搜尋 + 語意搜尋，撈出候選段落

語言模型列關鍵詞

③ 重排

逐段打分，挑出最相關的 6 頁；最高分低於 0.85 就回答「查不到」

重排模型（只會打分）

④ 抄寫核對

只能從這 6 頁逐字抄引句，程式再逐句核對

語言模型 + 程式

輸出：答案，每一點都能點回原文那一頁

這是「進階 RAG」，再加上專業領域必需的逐字核對。

參加測試的模型

名稱	模型	在哪裡跑	規模（參數數量）
8B	Qwen3-8B	本機筆電	80 億
9B	Qwen3.5-9B	本機筆電	90 億
27B	Qwen3.8-27B	本機筆電	270 億
雲端	Claude Opus 5.5	服務商伺服器	未公開，遠大於本機
8B、9B 微調版	同上，再訓練「抄引句」	本機筆電	同上

「B」= billion（十億），指模型內部參數的數量，可粗略理解為「腦容量」。

本機三個模型都在一台 48 GB 記憶體的筆電上執行，資料不出門。

六組模型走同一套四步流程、答同一批題目，只換第 ② 步與第 ④ 步的語言模型。

測驗：27 題，所有模型答同一批

題型	題數	例題
數值與參數	9	Hattian-Bala 壩高與湖水體積；唐家山洩流渠斷面
方法與方案	6	天然壩安定性快速評估；虹吸排水適用條件
機制、比較	5	有無溢洪道對潰決洪峰的影響
推理判斷	1	剛形成的堰塞湖能否給確切潰決時間？（正解：不能）
本來就查無	**6**	**南極洲、智利的堰塞湖案例；抽水機油耗**

每題都有直接從原文頁面整理的參考答案，逐題附原頁圖。

橫跨的範圍

答案分布在 23 份文獻：英文 9、繁中 8、簡中 6、日文 0

台灣、中國、巴基斯坦案例與跨國統計；文獻 1985–2026 年

今天唯一的比較基準

六個模型都回答這 27 題；之後所有數字都來自這 27 題

判分：五種結果

結果	定義	對使用者的影響
完整	每個檢查點都答到，而且沒有錯	可直接使用
部分	答到的都對，但漏了檢查點	需要自己補查
錯誤	任何一點錯（含引文正確但對象錯置、概念混用）	最危險：看起來有出處
誤判查無	知識庫有答案，卻回答「查不到」	漏掉關鍵文獻
正確查無	知識庫沒有答案，也誠實回答「查不到」	不被誤導

檢查點 = 參考答案拆出的必答要點；第 20 頁用一題實例說明。

把文獻整理成帶頁碼的小塊

106 份文獻

2,805 頁

36,123 個小塊，每塊都標「哪份文獻、第幾頁」

段落 22,259 · 表格的一列 13,193 · 整張表 153 · 圖中文字 518

拆表格的一列可能會失去前後文

本題的答案在表格裡，拆出來只剩「Height of landslide dam：130 m」，沒寫是哪座壩。搜尋時很難把它對上「Hattian-Bala」。

[Kiyota 2011｜第 3 頁｜表：Hattian-Bala 堰塞壩
Hattian-Bala 堰塞壩｜壩高（Height of landslide dam）：130 m

關鍵詞搜尋與語意搜尋一起用

關鍵詞搜尋：比對字面

精準；但中文問題找不到英文論文

語意搜尋：比對意思

中英互通；但比較粗，又稱「模糊搜尋」

文獻語言	系統怎麼處理	27 題中正確頁送到
繁中（43 份）	關鍵詞直接比對	9 / 9
簡中（12 份）	關鍵詞自動轉成繁、簡兩種寫法	6 / 6
英文（43 份）	另列英文關鍵詞 + 整題英譯	7 / 7
日文（8 份）	只靠語意搜尋（多語模型）	沒有題目，未驗證

例：7 / 7 = 答案在英文文獻的 7 題，正確頁都送到了模型手上

Hattian-Bala 這題撈到什麼

「巴基斯坦 Hattian-Bala 堰塞壩的壩高與湖水體積是多少？」

模型列的關鍵詞：Hattian-Bala · 堰塞壩 · 壩高 · 湖水體積 · 巴基斯坦 · landslide dam · dam height · lake volume · 另加整題英譯

搜尋方式	用什麼查	正確頁（Kiyota 2011 第 3 頁）排第幾
關鍵詞	8 個詞（繁簡兩種寫法），取前 12 頁	第 1 名（原文有 Hattian-Bala）
語意（中文題）	中文整題的意思，取前 40 段	沒進前 40
語意（英譯題）	英譯整題的意思，取前 40 段	第 1 名

合併後 **178 段、57 頁**：只有語意找到 45 頁、只有關鍵詞 8 頁、兩邊都找到 4 頁（正確頁在這 4 頁中）

兩邊不是投票：**候選全部合在一起，交給下一步的重排模型重新打分**

程式怎麼從 2,805 頁挑出這 6 頁

1 **關鍵詞找頁**：8 個詞在全庫比對，越罕見的詞分數越重；出現在超過 5% 頁面的泛詞（如「堰塞湖」）不能單獨讓一頁入選 → 取前 12 頁

2 **語意找段**：中文題、英譯題各取意思最接近的 40 段

3 **合併**：178 段、分屬 57 頁

4 **重排打分**：每頁取最高分那段當頁分數，57 頁排名。

5 **取前 6 頁**，再套右邊三條規則

6 **送出**：每頁最多 4,500 字（高分段落優先

規則 A　案例保障

題目點名的案例，提到它的頁至少要有一頁在 6 頁內；沒有就換掉最後一頁。本題 Kiyota 第 3 頁已在第 6 名，不必換。

規則 B　標題保障

題目點名的作者或案例，如果只出現在少數幾份文獻的標題，就直接送那份文獻最相關的一頁。

規則 C　查不到門檻

第 1 名低於 0.85 → 回答「查不到」；但若符合規則 B，仍送那份文獻的頁，引句要通過核對才算數。

重排的分數怎麼算出來

1　切段：候選頁的原文切成段落；超過 700 字的段落再切成互相重疊的小段

2　配對：每一段分別和「中文原題」「英譯題」各配成一對

3　打分：重排模型把「題目 + 段落」一起讀，輸出 0 到 1 的分數，越接近 1 越像在回答這題；兩個題目取較高分

4　排頁：一頁的分數 = 頁裡最高分的那一段；所有候選頁依分數排序，送前 6 頁。第 1 名都低於 0.85 → 回答「查不到」

本題的 Kiyota 第 3 頁

41 段 × 2 個題目 → 每段兩個分數

最高的是公式說明那段（英譯題）0.978

→ 這頁分數 0.978，在 57 個候選頁中排第 6 → 送出

第 1 名唐家山頁 0.996 ≥ 0.85 → 不回答「查不到」



分數是「相關程度」，不是「答案正確的機率」

分數由模型依訓練經驗判斷「像不像在回答」；我們決定的是送什麼進去、門檻與送幾頁。

同一頁裡，不同段落的分數差很多

用上一頁的算法替本題的段落打分（每段：中文題一個分數、英譯題一個分數，取較高者）：

原文段落（節錄）	出處	中文題	英譯題	採用
「坝高82~124 m，体积2 037万 m³」	唐家山（別座壩）	0.996	0.166	0.996
「H is the maximum crest height… V is the volume…」（公式說明）	Kiyota 第 3 頁	0.895	0.978	0.978
表格列「Volume：62 million m³」	Kiyota 第 3 頁	0.038	0.651	0.651
表格列「Height of landslide dam…：130 m」（答案）	Kiyota 第 3 頁	0.052	0.500	0.500

答案那一列只拿 **0.50**：表格列沒寫是哪座壩。這頁能進前 6，靠的是公式說明的 0.978。

逐段打分，再取那一頁裡「最高的那一段」當整頁分數。

送出的前 6 頁：模糊搜尋的優點與陷阱

名次	文獻	分數	誰找到	內容
1	唐家山堰塞湖應急處置（人民長江 2008）	0.996	語意	唐家山壩高、體積
2	唐家山滑坡與堰塞壩工程地質（人民長江 2008）	0.992	語意	唐家山壩高、體積
3	馬太鞍溪堰塞湖調查（2026）	0.988	語意	馬太鞍壩高、體積
4	既有堰塞湖處置案例（農村水保署）	0.987	語意	台灣某湖壩高
5	唐家山堰塞湖應急處置與減災（中國工程科學）	0.985	語意	唐家山體積
6	Kiyota et al. 2011 · Hattian-Bala（第 3 頁）	0.978	關鍵詞 + 語意	正確答案表格

懂意思、跨語言，但不在乎「是哪一座壩」：關鍵詞排第 1 的正確頁，重排後掉到第 6，剛好擠進送出範圍。

只能逐字抄，程式再核對

模型不准自己寫數字，只能從送來的原文逐字抄句子；對不上原文的不採用；白話摘要出現引句裡沒有的數字，也會被刪。**這 27 題，六個模型共抄了 463 句：25 句對不上原文被擋下，438 句通過。**

同一題，9B 模型的兩張證據卡：

擋得住：摘要數字被刪

抄的原文：「Volume：62 million m³」（Kiyota 2011，第 3 頁）✓

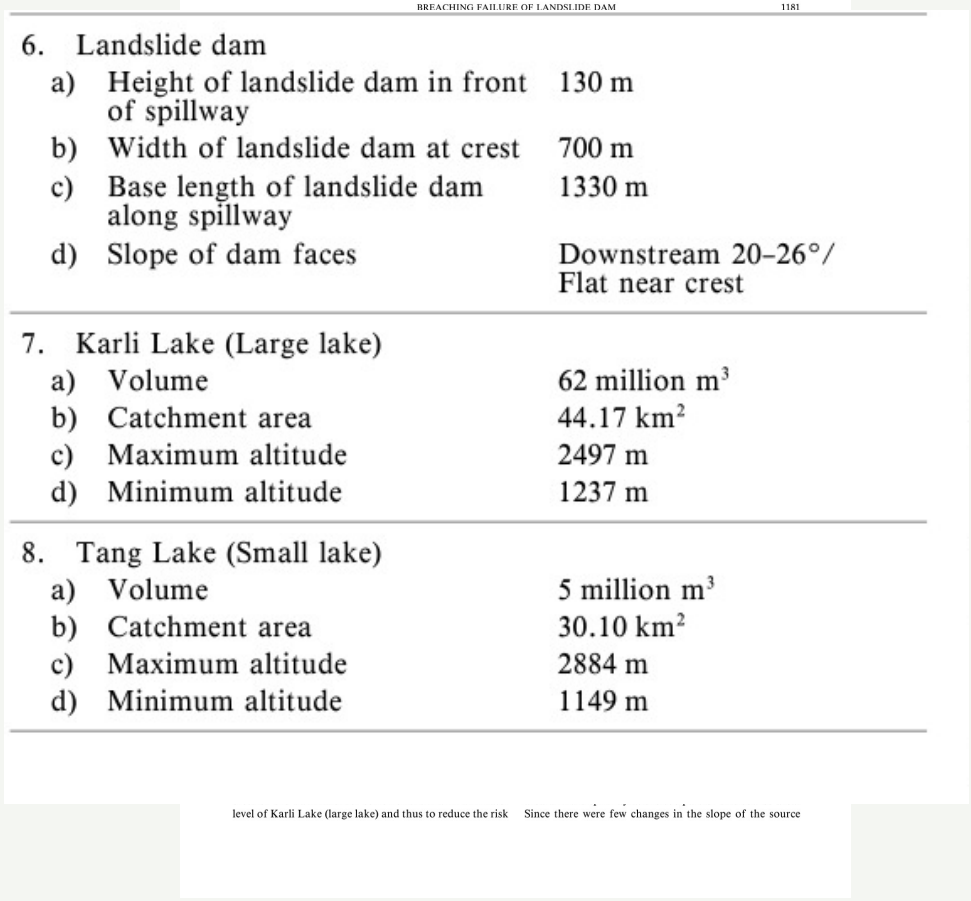
模型的白話：「6200 萬 m³」，寫法和引句不同 → **白話被刪除，只留原文**

擋不住：抄對句子、配錯對象

子問題「壩高」抄的原文：「壩高 82~124 m」（唐家山，人民長江 2008，第 1 頁）✓ 通過核對

問題：這句講的是唐家山，不是 Hattian-Bala。程式只核對「句子是不是真的」，不核對「是不是在講題目那座壩」。

把參考答案拆成檢查點



BREACHING FAILURE OF LANDSLIDE DAM 1181

6.	Landslide dam	
a)	Height of landslide dam in front of spillway	130 m
b)	Width of landslide dam at crest	700 m
c)	Base length of landslide dam along spillway	1330 m
d)	Slope of dam faces	Downstream 20–26°/ Flat near crest
7.	Karli Lake (Large lake)	
a)	Volume	62 million m^3
b)	Catchment area	44.17 km^2
c)	Maximum altitude	2497 m
d)	Minimum altitude	1237 m
8.	Tang Lake (Small lake)	
a)	Volume	5 million m^3
b)	Catchment area	30.10 km^2
c)	Maximum altitude	2884 m
d)	Minimum altitude	1149 m

level of Karli Lake (large lake) and thus to reduce the risk Since there were few changes in the slope of the source

① 壩高 130 m（溢洪道前方）

② 大湖 Karli Lake 62 million m^3

③ 小湖 Tang Lake 5 million m^3

④ 沒有錯置：不把同表的 250 m（堆積物最大厚度）當壩高，不混入別座壩

條件	結果
①②③ 都答到，④ 沒違反	完整
④ 沒違反，但少了 ①②③ 任一點	部分
違反 ④（任何一句錯置）	錯誤，不論其他多完整
有答案卻說「查不到」	誤判查無

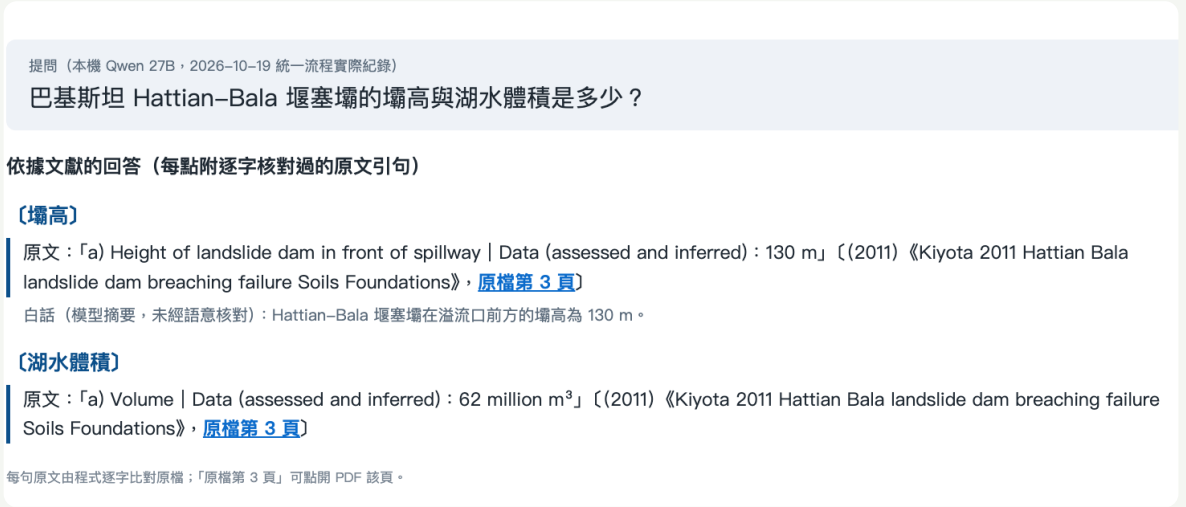
同一題、三個模型、三種結果

檢查點	雲端	本機 27B	本機 9B
① 壩高 130 m	有	有	有
② 大湖 62 million	有	有	有
③ 小湖 5 million	有	漏	有
④ 沒有錯置	有	有	混入唐家山
結果	完整	部分	錯誤

9B 混入唐家山的「壩高 82~124 m」「容積 3.16 億 m³」：引文是真的、核對也通過。

引文是真的，答案卻是錯的。

本機 27B 的實際答案畫面



每點的「原檔第 3 頁」可點開 PDF，就是第 20 頁那張原頁；漏掉的小湖就在同一張表的下一列。

04 整體結果

每一步的取捨，都有實際例子

步驟	實際發生的例子	限制
① 切塊	Hattian-Bala 的答案列被切出來後沒有壩名，重排只得 0.50（第 17 頁）	補上表格標題能救。
② 語意搜尋	Hattian-Bala 題，重排後前 5 名都是別座壩的「壩高、體積」（第 18頁）	意思相近不等於有答案
③ 重排	Hattian-Bala 正確頁只排第 6 名	只送前 6 頁，第 7 名以後模型看不到
③ 門檻	27 題中，有答案的題最高分最低 0.886；沒答案的題最高分最高 0.840，門檻 0.85 夾在中間	間距只有 0.046，換一批題目可能被跨過
④ 核對	463 句引句中 25 句對不上原文被擋下；摘要「6200 萬」被刪（第 19 頁）	擋不住對象錯置（第 19 頁）

每一步漏掉的，下一步都補不回來。

兩種「查不到」：誤判與正確

誤判查無（8B）

題目：唐家山洩流渠的設計斷面尺寸與渠底高程？

正確頁面已送到，模型卻回答「內容不足以回答」。

答案就在它看過的頁上：高標準 742 m / 13 m、中標準 745 m / 22 m、低標準 747 m / 28 m，邊坡 1:1.5。

正確查無

題目：知識庫裡有沒有南極洲的堰塞湖案例？

系統回答「查無足以回答的資料」，並列出比對的關鍵詞（南極洲、Antarctica…）與查過的範圍。

「查不到」要附上查了什麼、看了哪些頁，使用者才分得出是真的沒有，還是系統漏掉。

AI 知不知道自己找不到？

6／6

本來就查不到的 6 題，六個模型全部答「查不到」

1–3 題

但有答案的 21 題裡，8B、9B 小模型各有 1–3 題誤判「查不到」

能做到，靠的是兩道程式防線：分數不到門檻就不送資料；引文對不上原文就不採用。

六個模型，同一套流程、同一批題目

模型	位置	完整	部分	錯誤	誤判查無	每題秒數
Claude Opus 5.5	雲端	21	0	0	0	約 45
Qwen 27B	本機	17	4	0	0	134
Qwen 9B 微調	本機	10	9	1	1	54
Qwen 8B 微調	本機	10	4	2	3	65
Qwen 8B	本機	9	8	1	3	79
Qwen 9B	本機	7	9	3	2	67

可答 21 題；另 6 題查無題，六組全部答對

怎麼讀這張表

1 本機 27B 做到 **0 錯誤**，與雲端差 4 題，差在「抄得不夠齊」，不是答錯。

2 8B、9B 落在 7–10 題，差距在重跑浮動內（同題重跑可能差 1–2 題）。

3 錯誤都是「對象錯置」或「概念混用」，**沒有一題是編造數字**。

04 整體結果

① 整理 程式＋校正 → ② 搜尋 語言模型列關鍵詞 → ③ 重排 重排模型（只會打分） → ④ 抄寫核對 語言模型＋程式

小模型能靠微調追上嗎？

次	模型	改哪一步	訓練資料	結果（題數）
1	8B	④ 抄引句	95 筆示範	完整 ＋1 · 錯誤 −3
2	9B	④ 抄引句	95 筆示範	完整 ＋3 · 錯誤 −2 · 誤判查無 −1
3	9B	② 關鍵詞	224 題	正確頁送到 ＋1
4	重排模型	③ 排名	3,762 組題目–段落	正確頁送到 ＋8

看得到的改善跡象

8B 微調後錯誤 4 → 1–2；六模型比較中，9B 微調版錯誤 3 → 1

為什麼還不能說有效

四次都沒達到事先定好的標準；資料只有 95–3,762 筆；差距多在重跑浮動（±1–2 題）內

04 整體結果

微調的初步結論（限本研究的四次嘗試）

改「抄引句」的兩次：完整題增減都在 2 題以內，與重跑浮動同一量級，**看不出穩定的改善**。

改「排名」的一次：正確頁送到的題數多了8題。

以目前的資料量與題數，**既不能說微調有效，也不能說微調無效**。需要更大的訓練集、另一批沒看過的題目，並事先定好驗收標準。

能確定的是：瓶頸在「找對頁」。

05 大模型一定比較好嗎？

評估指標

調查效率

每題約 2 分鐘

本機 27B；雲端約 45 秒

資料精度

完整 17 / 21

本機 27B，錯誤 0 題

漏檢率

21 / 21 送到

有的只排第 6、靠規則補送

引用可靠度

擋下 25 句

463 句引句，438 句通過核對

技術成熟度

重跑 ±1–2 題

溫度設 0 仍會浮動

設備成本

一台 48 GB 筆電

可跑 27B；微調借雲端 GPU

作業風險

對象錯置

編造引文已擋得住，這項還沒

實務可行性

全程本機

資料更新可比對、可回復

成本與速度：本機 vs 雲端

本機 27B

- 資料不出門
- 不按次數付費
- 每題約 2 分鐘
- 需要高記憶體硬體

雲端

- 最準、最快
- 文獻片段要送出去
- 需要授權
- 服務商保留一段時間，無法收回

機關文件、未公開調查資料不一定能上雲端：本機 27B 少 4 題完整，但一樣沒有錯誤。

風險：防得住的與防不住的

防得住

- 編造引文（逐字核對）
- 知識庫沒有的題硬答（門檻）
- 無法核對出處（每點可點回原頁）

防不住

- 引文對、對象錯（第 21 頁的唐家山數據）
- 正確頁沒送到而誤判查無
- 文獻本身的矛盾或過時資料

定位：幫專家快速找到可核對的證據；結論由專家下。

一定要用最大的模型嗎？

1 準確度上，大模型確實比較好：雲端 21 > 本機 27B 17 > 小模型 7–10。

2 但本機 27B 已經「可用」：0 錯誤、每點可核對、資料不出門。

3 四次嘗試中，小模型加微調沒有追上 27B；資源放在「找對頁」比微調寫答案更值得。

選模型不只看分數：資料能不能送出去、可以等多久、錯一題的代價多高。

AI 時代：模型會換，資料會留下

模型 = 路上的車

每幾個月

就有新一代，隨時可以換

資料庫 = 鋪好的路

數十年

的調查與文獻，整理好就能一直用

路的規格：答案 → 證據 → 文獻 → 頁碼 → 原始檔案

「對的資料」：每個數字都要帶著身分

1　對象

哪一座壩、哪一次事件

2　性質

設計值、實測值，還是模擬值

3　條件

在什麼水位、什麼時間點

4　名稱

中英日名稱、舊地名、不同拼法

這幾個欄位定好了，換什麼模型都用得上。

為什麼需要各位領域專家

最難保存的是判斷

文獻之間會互相矛盾

數字會有歧義

現場經驗多半沒寫下來

AI 能把它們並列，
哪個才對，只有懂的人能判斷。

專家可以做的四件事

1. 提出實務上真的會問的問題
2. 定義怎樣算答對
3. 裁決文獻之間的衝突
4. 決定保存什麼、怎麼標記

最大的阻力：每個人對 AI 的認知差距很大

計畫方、審查委員、第一線使用者、資訊人員，期待常落在光譜的兩端

過度期待

「接上資料庫就好」
→ 資料沒整理好，AI 一樣會配錯對象

「微調一次就會變好」
→ 要持續累積資料、持續驗證

全盤否定

「AI 會亂編，不能用」
→ 可要求每點附原頁，由程式逐字核對

「一定要用最貴的雲端」
→ 依資料敏感度與錯誤代價來選

兩端都需要同一套看得到的衡量方式。

06 展望

縮小落差：用同一套看得到的標準

1

分開報結果

不只一個準確率，五種結果分開計

2

先定標準，再看結果

避免看了結果再改標準

3

每一點都能點開

不必相信 AI，自己查原頁

驗收不寫「準確率 90%」，改寫：題目從哪來、誰定答案、錯誤最多幾題

06 展望

往下鋪路：地質 AI 基礎建設的三個階段

近期

既有文獻
變成可追溯的知識庫

每個答案都能點回原頁

中期

跨單位共用同一套標記

描述欄位、名詞對照一起定

長期

串起調查、監測、
處置紀錄與法規

事件之間互相連結，可以交叉查證

越往後，越需要跨單位一起定規則。

06 展望

回到研究問題：可以，但有條件

它是專家的查資料助理，不是決策者。

資料先整理好

比選模型更重要

專家定義「對」

事先定好驗收標準

每點都能點回原頁

並保留「查不到」的權利

什麼證據值得保存，要靠在座的各位。