TASPAA 2025 國際學術研討會
自組場次：AI衝擊下的數位治理演進

RAG知識管理系統之POC實驗評估

廖洲棚 副教授
國立空中大學公共行政學系

分享大綱

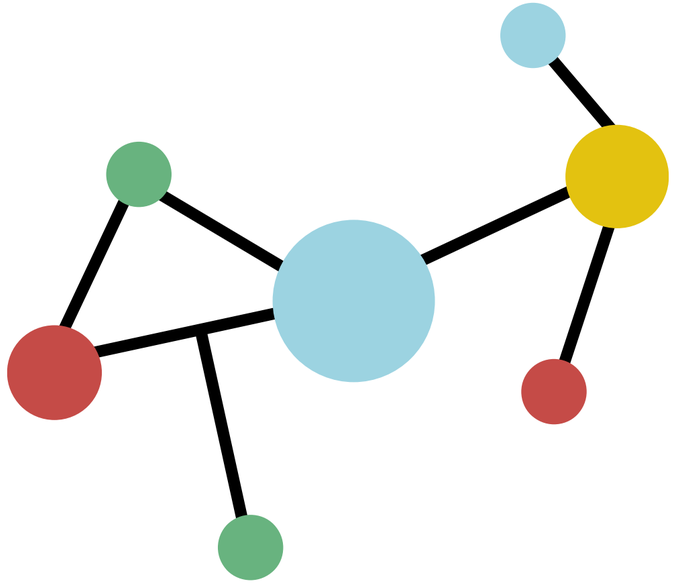
壹、研究背景

貳、POC實驗評估設計

參、POC實驗評估結果

肆、小結

壹、研究背景



壹、研究背景

生成式人工智慧（Generative Artificial Intelligence, GenAI）的崛起

- 資通訊科技的發展：生成式AI技術迅速崛起，廣泛應用於提升行政效率和支援業務流程。
- 政府數位轉型的需求：在數位化轉型的背景下，各國政府需有效運用生成式AI以提升治理效能。

公部門運用GenAI的挑戰

- 數據可靠性：GenAI會出現為人詬病的「幻覺」（hallucination），提供貌似合理但實際錯誤的訊息。
- 建置成本與技術門檻高：大型語言模型（Large Language Model, LLM）的建置成本及技術門檻皆高，導致政府面臨掌控權不足的問題。
- 資訊安全風險：政府機關的機敏性資料基於資訊安全風險問題，不能與商業的LLM分享。

壹、研究背景

RAG系統的優點：

1. 使用者可以使用自然語言和系統對話。
2. 使用自然語言處理技術，建置向量文件資料庫，克服內部資料暴露於外的風險並提高檢索效能。
3. 結合大型語言模型，降低系統建置成本及增強內容生成能力降低幻覺問題。

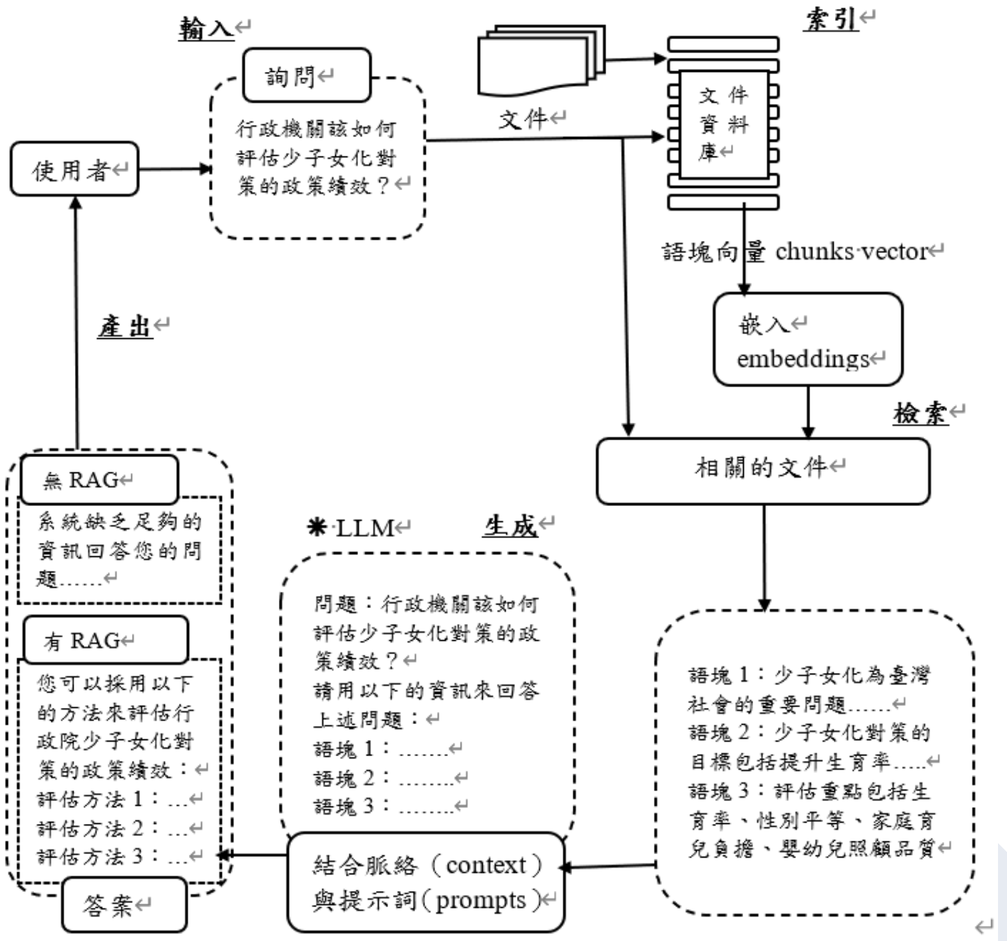
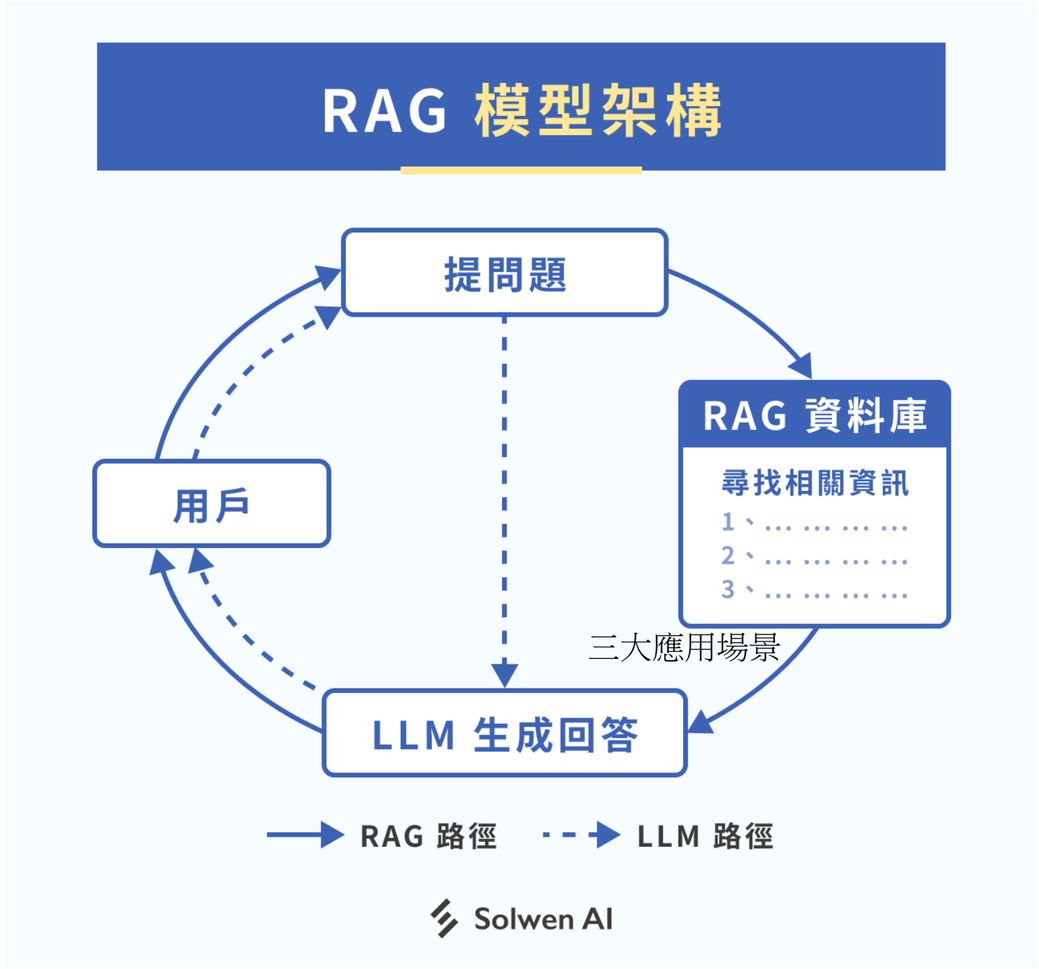


圖1 RAG技術的運作流程

資料來源：修改自Gao et al. (2023)。



三大應用場景

- **企業知識管理**：讓企業內部人員透過自然對話，輕鬆找到需要的正確資訊
- **客服機器人**：24 小時快速精準回覆客戶問題，提升客戶滿意度
- **專業知識問答**：打造創新服務模式，擴大客戶觸及範圍、降低營運成本

資料來源：https://solwen.ai/posts/what-is-rag

壹、研究背景

RAG系統運作原理：

1. 利用自然語言處理文件，建立embedding，資料庫同時儲存文件向量及文件。
2. 使用者使用自然語言詢問問題，系統將問題轉成embedding，透過比對這個問題的向量與文件向量的相似度，找出最相關的幾篇資料。
3. 把找到的相關文件的最相關段落轉成語塊（chunks）傳給介接的大型語言模型（如ChatGPT-4o），模型根據檢索來的資料和使用者問題生成答案。

	Chunk Index	Cosine Similarity
4	4	0.8766172429165234
3	3	0.803560521547084
0	0	0.7729587700723243
1	1	0.6699097202659685
2	2	0.522803854398557

	Chunk
1	生成式人工智慧（Generative Artificial Intelligence, GenAI）的崛起引起公共管理 Gao等人（2023）將RAG
2	強生成（RAG）技術，似乎為前述成本及風險問題的解決帶來了新的曙光。 Gao等人（2023）將RAG技術的運作流程描繪如圖1所示，以RAG技術建置的系統
3	在這個步驟中所有文件被分成語塊（chunk）且編碼為向量（vector）以儲存在向 作者以圖1所示的實例進行解說，方便讀者理解RAG系統的運作流程。首先，行政
4	實例進行解說，方便讀者理解RAG系統的運作流程。首先，行政機關可以將與少
5	題以及系統檢索出的語塊，組合成適合問題情境的提示語（prompts）並向LLM提 由於RAG系統的使用者可以直接使用自然語言和系統對話，並因此迅速地協助使

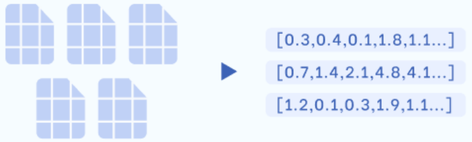
RAG 在企業中的應用流程

索引 將企業的資料經過預處理放進資料庫中

1.文檔加載

1.文檔分割

3.嵌入



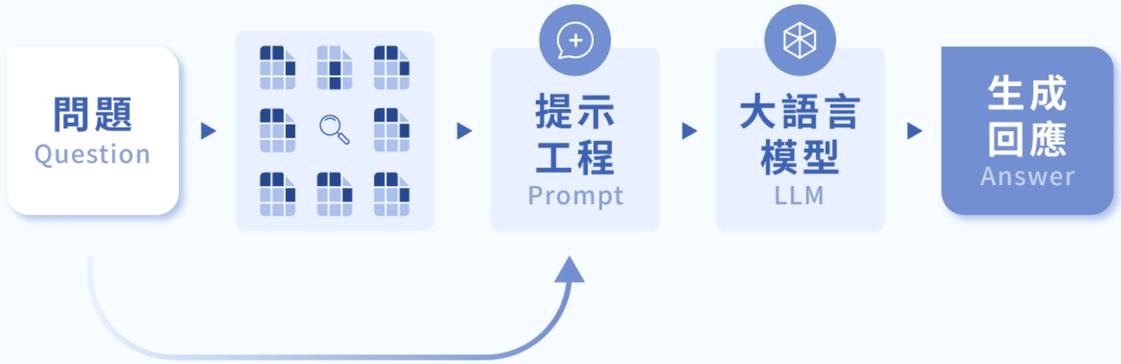
4.存儲

Solwen AI

RAG 在企業中的應用流程

檢索和生成

- 當使用者做詢問時，根據提示詞通過檢索演算法找到最接近的幾段資料。
- 把這幾段資料也納入提示詞，讓生成式AI可以根據這些相關資料來做回覆，提升回答的準確性。



Solwen AI

資料來源：https://solwen.ai/posts/what-is-rag

壹、研究背景

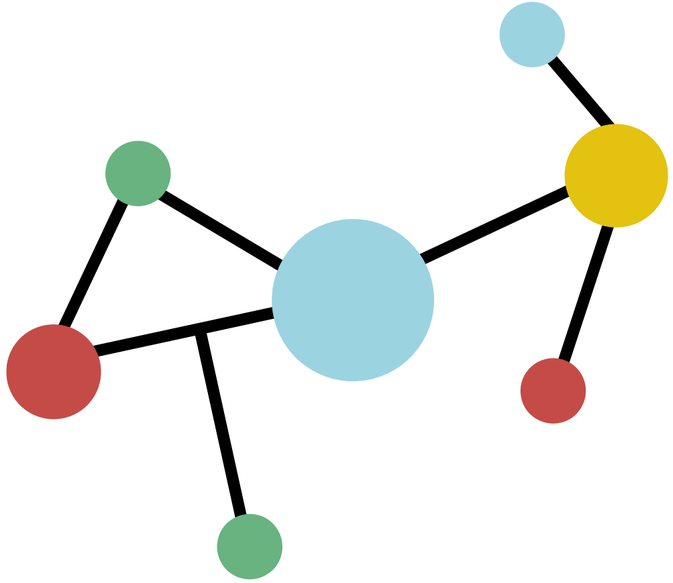
RAG系統的優點：準確性、即時性、低成本、客製化

1. 使用者可以使用自然語言和系統對話。
2. 使用自然語言處理技術，建置向量文件資料庫，克服內部資料暴露於外的風險並提高檢索效能。
3. 結合大型語言模型，降低系統建置成本及增強內容生成能力降低幻覺問題。
4. 可依據組織的需要開發適合的RAG系統。

壹、研究背景

- 需要建構一套用於系統概念驗證（Proof of Concept, POC）之實驗評估方法，協助公部門驗證運用檢索增強生成（Retrieval Augmented Generation, RAG）技術開發知識管理資訊系統的效能。
- POC實驗評估的重點：
  1. 檢索功能：「檢索模組能找出與問題相關的內部文件」，並能「聚焦且淬取出與問題相關的內容段落（context passages）」以供給LLM。
  2. 生成功能：「LLM能忠實地利用這些內容段落」，並「生成能正確回答使用者提問」的「高品質答案」。

貳、POC實驗評估設計



貳、POC實驗評估設計

實驗設計

- 參與對象：「實際負責規劃或執行之資深同仁或科長級以上主管」公務人員。以實機測試加上網路問卷調查受試者評估意見的方式進行四次測試。

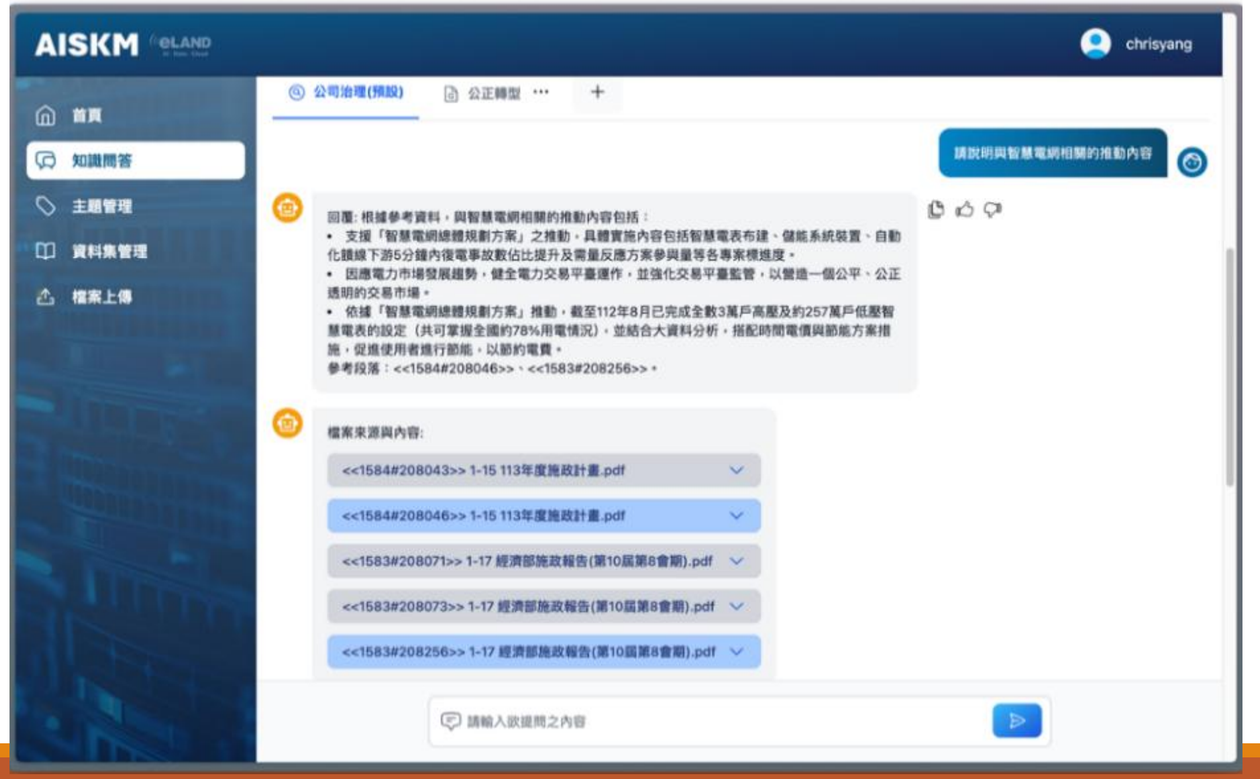
評估指標

- 檢索功能：上下文召回率（Context Recall）、上下文相關性（Context Relevance）等兩項指標。
- 生成功能：答案相關性（Answer Relevance）、忠實度（Faithfulness）、答案正確性（Answer Correctness）等三項指標。
- 使用者整體滿意度指標。

貳、POC實驗評估設計

- 以意藍資訊開發的「AI Search for KM」系統做實驗評估平台。

平台服務畫面



貳、POC實驗評估設計

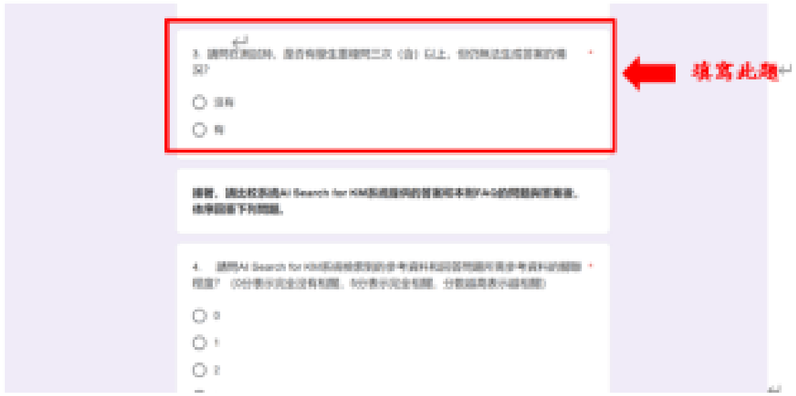
評估構面	評估指標	人工評估定義	自動評估定義	調查題目
系統 檢索功能	上下文召回率 (Context Recall)	檢索出的文本與標準答案的關聯程度	檢索出的段落與標準答案之關聯性。（0～1分）	請問AI Search for KM系統檢索到的參考資料和回答問題所需參考資料的關聯程度？（0～5尺度）
	上下文相關性 (Context Relevance)	檢索內容與問題的相關程度	檢索出的段落與問題之關聯性。（0～1分）	請問AI Search for KM系統檢索出的參考資料和這則FAQ問題的相關程度？（0～5尺度）
系統 生成功能	答案相關性 (Answer Relevance)	系統回答與使用者提問的相關程度	評估生成的回應與問題之關聯性。（0～1分）	請問AI Search for KM系統提供的答案和這則FAQ問題的相關程度？（0～5尺度）
	忠實度 (Faithfulness)	系統回答能否忠實反映檢索出的文本內容	評估生成的回應與檢索段落之關聯性。（0～1分）	請問AI Search for KM系統能依據參考資料提供答案而非胡言亂語的程度？（0～5尺度）
	答案正確性 (Answer Correctness)	系統回答與標準答案一致的程度	無	請問AI Search for KM系統提供的答案和這則FAQ答案的內容一致性程度？（0～5尺度）
整體效能	系統服務滿意度	系統提供服務的滿意度	無	整體而言，我對AI Search for KM系統提供的服務感到滿意？（1～5尺度）

貳、POC實驗評估設計

實驗指引守則

（二）填寫問卷的第 3 題題目「請問在測試時，是否有發生重複問三次（含）以上，但仍無法生成答案的情況？」。

1. 請將「FAQ 的題目」貼在 AI Search for KM 的對話框中，點選送出。**由於系統尚在發展，因此，若詢問系統時，無法看到生成的回應，請再點一次嘗試，若重複問三次（含）以上，但仍無法生成答案的情況，請在本題勾選「有」；若沒有發生無法生成答案的情況，或是在重複詢問後有生成答案（不管問幾次最終有生成答案皆可），請在本題勾選「沒有」。**



FAQ出題指引與範例

附錄三、FAQ 出題指引與範例

113 年度國家發展委員會「建置以資料科學為基礎之社會政策治理機制」委外服務計畫

計畫主持人：陳敦源教授（政治大學公共行政學系）
協同主持人：廖洲棚副教授（國立空中大學公共行政學系）
黃妍甄助理教授（暨南國際大學公共行政與政策學系）

常見問答集（FAQ）出題說明：

1. 本團隊為執行國家發展委員會「113年建置以資料科學為基礎之社會政策治理機制」委外服務計畫，評估社會政策智慧治理運用生成式 AI 之協作能力，規劃發一個輔助政策規劃的生成式 AI 系統，提供知識搜尋與問答功能，供公務同仁做為政策議題資料檢索及政策規劃之參考，非常感謝您撥冗，在此向您致上最高謝意！
2. 本期以「我國少子女化對策計畫(107至114年)」為研究範圍，除蒐集相關政策計畫及研究文獻外，邀請各機關熟悉相關業務人員共同建置問答集，透過高品質範例訓練 AI 模型，並進行實驗測試優化 AI 生成結果。
3. 請依據「FAQ 題號名單與題號對照表」找到貴機關業務所對應之題號，並從研究團隊提供的「FAQ 問答集」找到該題的回答，並針對實務實際情況來進行修正或補充，包含政策議題（或問題）、參考答案與參考資料等欄位皆可修正。亦或者您可根據以下範本的格式，提出更重要的政策議題（或問題），並撰寫新的參考答案（字數至少需撰寫800-1000字），以及列出參考答案所需的參考資料，來取代研究團隊的回答，相關內容請填寫至「FAQ 問答集表格」中。
4. 請於113年10月31日（星期四）之前，將修正後或新增的回答（word 檔）以及參考資料的電子檔（word/pdf 檔），寄至以下本計畫的專用信箱：pmo20240308@gmail.com，俾無修正或新增回答，以及補充參考資料者，亦請回信告知。
5. 上述參考資料若有《政府資訊公開法》第18條規定之情形者，則無需提供或僅就得公開部分提供。

感謝您的參與，若有任何問題，歡迎隨時與我們聯繫，聯絡資訊如下：

實驗者資料表

附錄六、實驗者資料表

113 年度國家發展委員會「建置以資料科學為基礎之社會政策治理機制」委外服務計畫
「社會發展政策知識庫」測試實驗說明及參與者邀請

計畫主持人：陳敦源教授（政治大學公共行政學系）
協同主持人：廖洲棚副教授（國立空中大學公共行政學系）
黃妍甄助理教授（暨南國際大學公共行政與政策學系）

壹、辦理「社會發展政策知識庫」測試實驗之目的

為驗證檢索增強生成（RAG）技術應用於政策知識庫的可行性，本計畫將建置「社會發展政策知識庫」做為前述概念驗證（POC）的標的。同時，為驗證該系統在公務場域應用的潛力，本研究團隊特規劃辦理本次測試實驗，以評估系統於輔助公務人員從事政策規劃相關工作的成效。本測試實驗規劃邀請有執行我國少子化對策計畫（107 年-114 年）經驗之公務同仁參與，參與者需配合本實驗的指引依序參與 4 次測試，並在每次測試後填寫調查問卷，回饋系統使用經驗，提出參考資料補充建議並提供建議補充之參考資料的電子檔，以協助研究團隊持續優化系統的功能。

貳、參與測試實驗者之權利與義務

本計畫誠摯地邀請有執行我國少子女化對策計畫（107 年-114 年）經驗之科長級主管同仁參與測試實驗並回饋系統使用經驗，茲就參與者之權利義務說明如下：

1. 請填寫完整的聯絡資訊，包含姓名、職稱、任職機關與單位名稱、聯絡電話、e-mail、通訊地址等資訊，以利研究團隊於測試實驗過程的聯繫以及商品禮卷的寄送。
2. 研究團隊將嚴格保密您的個人資訊，任何個人識別資訊將被去識別化處理，並於本研究完成後一年銷毀您的個人資訊。
3. 本測試實驗不涉及任何可能對您造成心理或生理負擔的問題，若您在參與過程中感到不適，可以隨時選擇退出。
4. 本計畫備有超商禮卷贈與參與者，以感謝參與者的專業協助，每參與 1 次實驗並填寫調查問卷，可獲得 300 元超商禮卷，完成 4 次實驗者，可獲得共計 1500 元之超商禮卷。中途退出測試實驗者，按實際參與次數贈與超商

貳、POC實驗評估設計

發展RAG系統評估量表，邀請公務人員協助建構問答集進行系統測試，評分檢索與生成功能。

113 年度國家發展委員會「建置以資料科學為基礎之社會政策治理機制」委外服務計畫

計畫主持人：陳敦源教授（國立政治大學公共行政學系）
協同主持人：廖洲棚副教授（國立空中大學公共行政學系）
黃妍甄助理教授（國立暨南國際大學公共行政與政策學系）

我國少子女化對策計畫常見問答集（FAQ）測試指引說明：

1. 以下為您的測試題目以及回答，包含您協助製作的題目與回答、參考資料。。
2. 請依照以下您專屬的帳號與密碼完成測試。測試步驟請參見「POC 實驗指引守則」。

單位：_____
受測者：_____
您的個人代碼：_____
帳號：_____
密碼：_____

感謝您的參與，若有任何問題，歡迎隨時與我們聯繫，聯絡資訊如下：
黃妍甄助理教授，電話：04-92910960 分機：2742
信箱：yen1227@mail.ncnu.edu.tw
簡翊昀研究助理，電話：0926662331
信箱：angelchien914@gmail.com

給實驗者的測試指引說明

測試邀請信件

貳、POC實驗評估設計

- 本研究除了針對RAG的檢索與生成功能進行評估外，在第四次實驗中，另加入用於評估對於本研究使用生成式AI模型ChatGPT-4o所生成的答案的題項，以比較RAG系統和生成式AI在生成答案上的差異。

參、POC實驗評估結果

參、POC實驗評估結果

各階段實驗執行情形

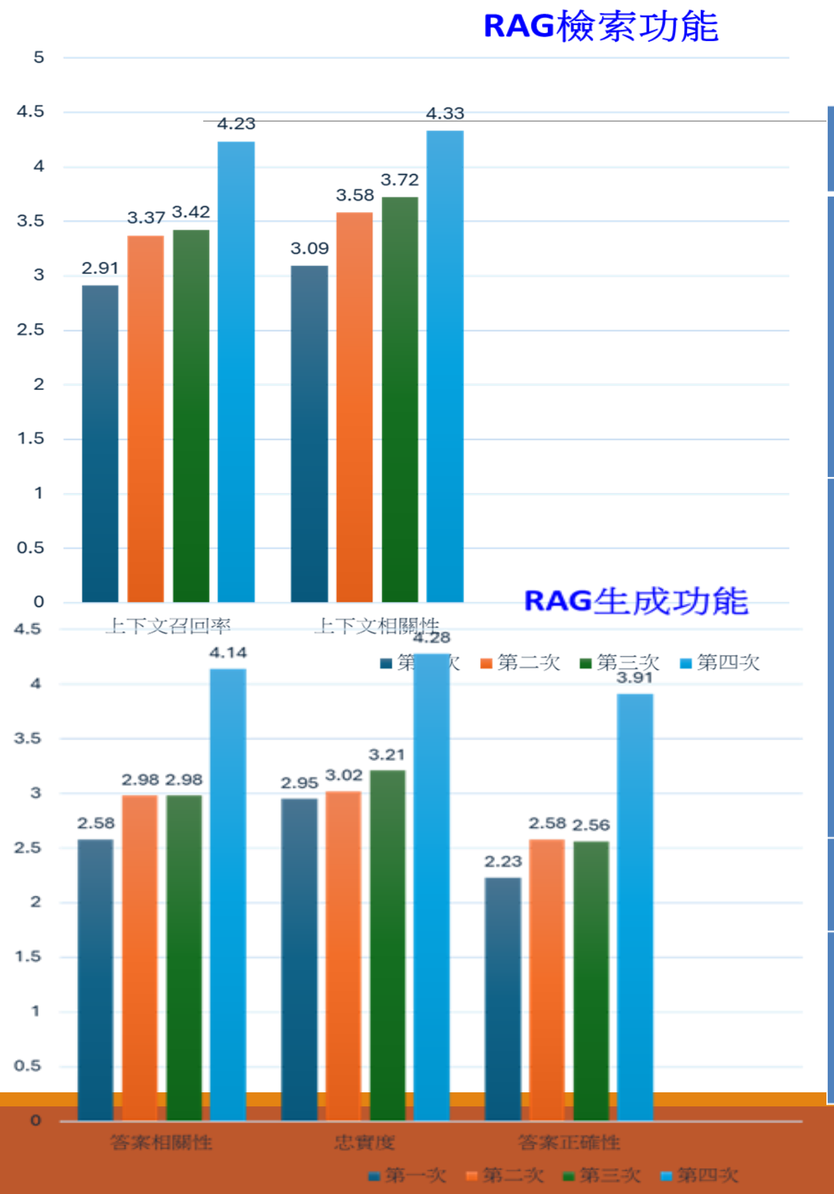
	第一次實驗	第二次實驗	第三次實驗	第四次實驗
時間	2024年11月18日至12月12日	2024年12月17日至12月24日	2025年01月10日至01月24日	2025年01月24日至02月07日
參與實驗人數	22	22	22	21
實驗題數	43	43	43	43
改善建議題數	7	6	7	7

參與實驗者的個人資料分布情形

變項	類別	百分比（個數）
性別	男性	57.1%（12）
	女性	42.9%（9）
年齡	20-29歲	4.8%（1）
	30-39歲	38.1%（8）
	40-49歲	42.9%（9）
	50-59歲	14.3%（3）
學歷	博士	9.5%（2）
	碩士	81.0%（17）
	學士	9.5%（2）
官等	無	28.6%（6）
	薦任	66.7%（14）
	簡任	4.8%（1）
職等	七等	9.5%（2）
	九等	42.9%（9）
	八等	19.0%（4）
	十等	9.5%（2）
	無	19.0%（4）

註：全程參與實驗評估人數：21人

參、POC實驗評估結果

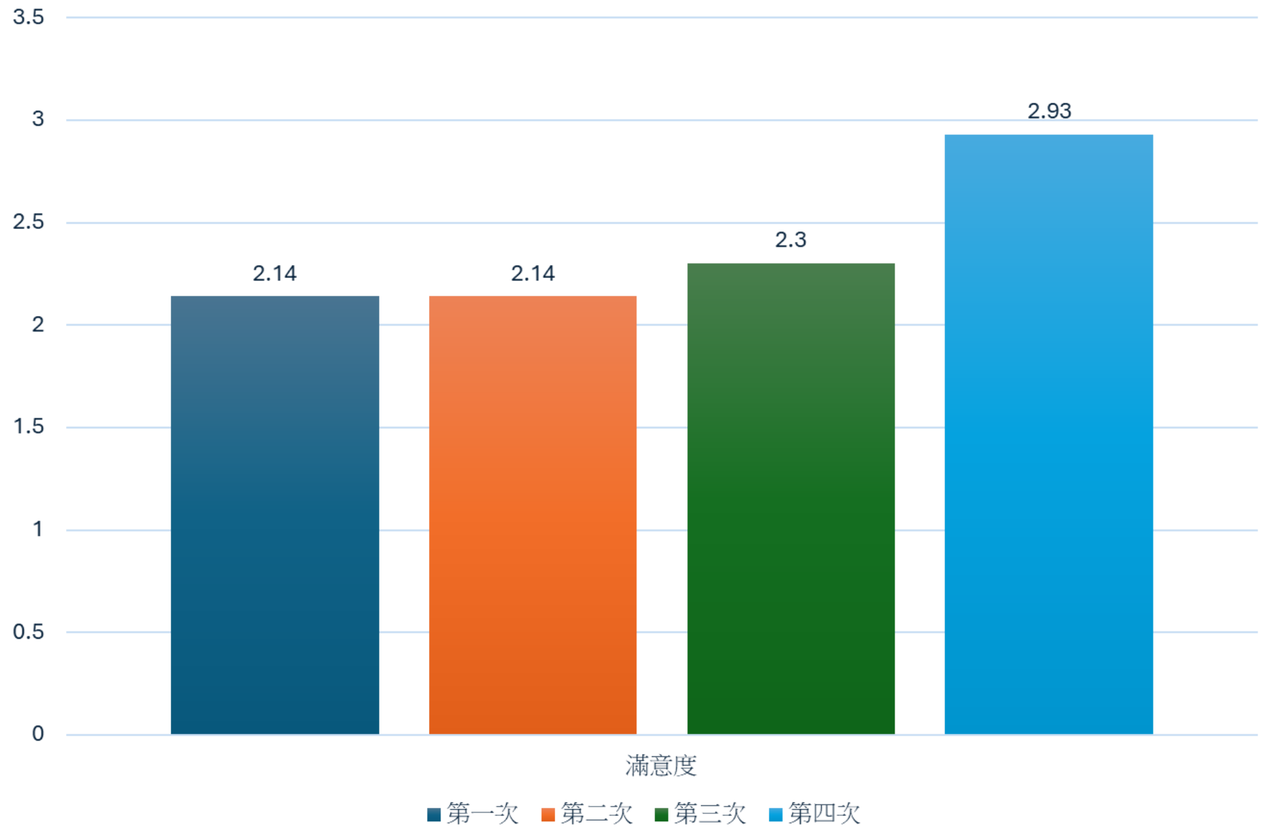


第一次與第四次實驗之人工評估結果比較

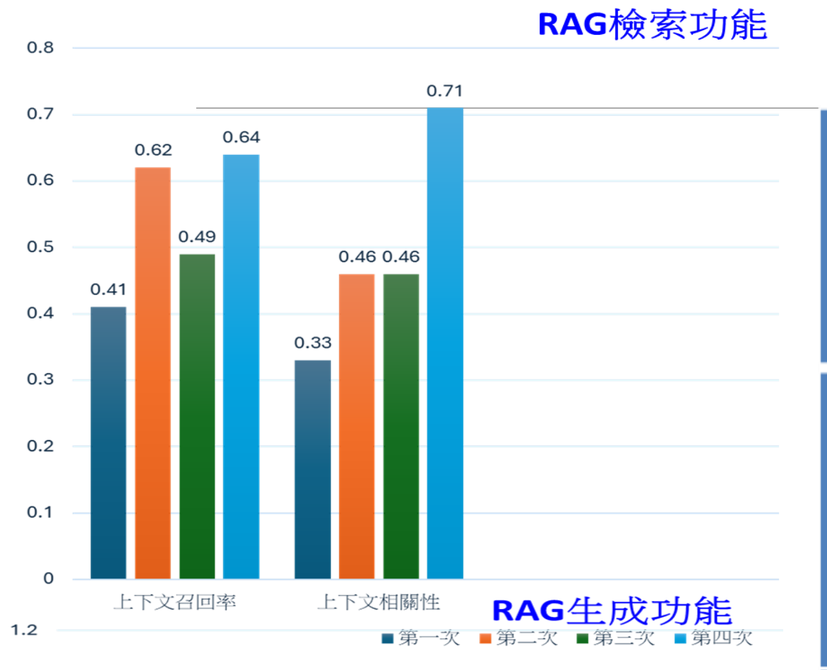
評估構面	指標	類別	平均分數	FAQ題數	標準差	T值	顯著程度
RAG檢索功能	上下文召回率	第一次實驗	2.91	43	1.41	-5.336	0.000
		第四次實驗	4.23	43	1.02		
	上下文相關性	第一次實驗	3.09	43	1.29	-5.531	0.000
		第四次實驗	4.33	43	0.81		
	檢索功能	第一次實驗	3.00	43	1.27	-6.016	0.000
		第四次實驗	4.28	43	.819		
RAG生成功能	答案相關性	第一次實驗	2.58	43	1.71	-5.506	0.000
		第四次實驗	4.14	43	0.83		
	忠實度	第一次實驗	2.95	43	1.56	-5.289	0.000
		第四次實驗	4.28	43	0.77		
	答案正確性	第一次實驗	2.23	43	1.63	-6.802	0.000
		第四次實驗	3.91	43	0.84		
	生成功能	第一次實驗	2.59	43	1.54	-6.224	0.000
		第四次實驗	4.11	43	0.74		
服務滿意度	系統服務滿意度	第一次實驗	2.14	43	0.97	-5.240	0.000
		第四次實驗	2.93	43	0.63		
總平均	總平均	第一次實驗	2.75	43	1.36	-6.537	0.000
		第四次實驗	4.18	43	0.72		

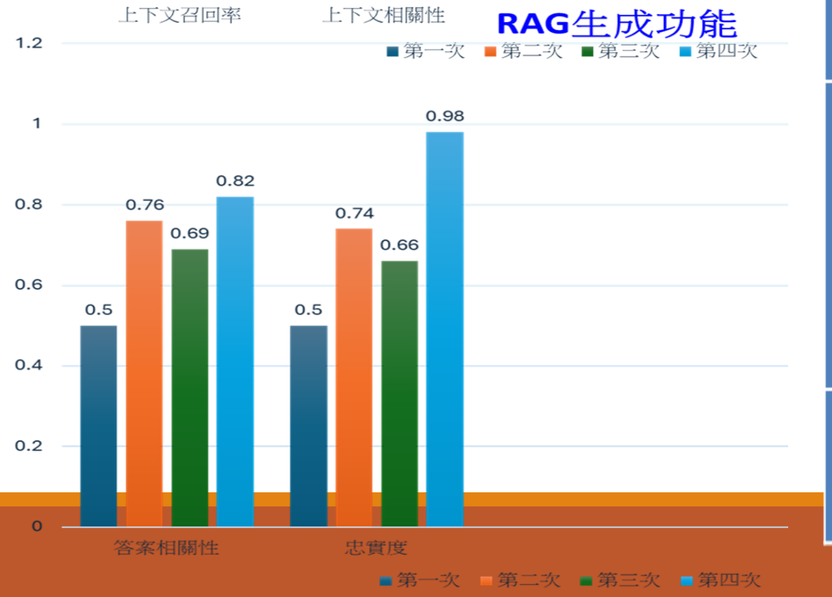
參、POC實驗評估結果

系統服務滿意度



參、POC實驗評估結果





第一次與第四次實驗之自動評估結果比較

評估構面	指標	類別	平均分數	FAQ題數	標準差	T值	顯著程度
RAG檢索功能	上下文召回率	第一次實驗	0.41	43	0.45	-3.234	0.002
		第四次實驗	0.64	43	0.27		
	上下文相關性	第一次實驗	0.33	43	0.36	-6.876	0.000
		第四次實驗	0.71	43	0.22		
RAG生成功能	答案相關性	第一次實驗	0.50	43	0.50	-4.405	0.000
		第四次實驗	0.82	43	0.12		
	忠實度	第一次實驗	0.50	43	0.50	-6.072	0.000
		第四次實驗	0.98	43	0.07		
總平均	總平均	第一次實驗	0.44	43	0.44	-5.467	0.000
		第四次實驗	0.79	43	0.93		

參、POC實驗評估結果

ChatGPT與RAG系統之生成功能人工評估結果比較

指標	類別	FAQ題數	標準差	標準差	T值	顯著程度
答案相關性	RAG系統	4.14	43	0.833	-0.236	0.814
	ChatGPT	4.19	43	1.139		
答案正確性	RAG系統	3.91	43	0.840	1.450	0.154
	ChatGPT	3.65	43	1.173		

說明:上述題目之選項均採0-5分之計算，呈現分數為所有測試結果的平均分數。

肆、小結

肆、小結

實驗評估研究發現

1. 提供可行的POC實驗評估方法
2. 驗證RAG系統的在公部門知識管理的應用潛力

肆、小結

未來應用建議

1. 持續推動行政機關的資料治理工作
2. 建構可信任的AI系統
3. 增進公務人員使用AI系統的能力與意願


thank you
danke
謝謝
ngiyabonga
teşekkur ederim
gracias
merci
igado
dziękuję
sukriya
kop khun krap
go raibh maith agat
terima kasih
grazie
arigatô
dakujem
mochchakkeran
tapadh leat
obriga
mauruuru
hvala
dankie
kiitos
dhanyavad
asибo
рахмат
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
mersi
barka
welalin
tack
djiere dieuf
tau
дякую
mamnun
chnorakaloutioun
gratias ago
gràcies
sulpáy
akun
dankon
ačiū
köszönöm
gràcie
bayarlalaa
nandri
nanmi
sobodi
dėkuji
mèsi
madloba
sagolun
najis tuke
ありがとう
tanemirt
rahmet
kam sah hamnida
ahmat
diolch
dhanyavadagalu
shukriya
мерс
xièxie
감사합니다
ευχαριστώ
takk
tạ ơn
tạiku