# 摸透語言模型的習性：LLM 會偏袒什麼樣的文章？
## —— 探討 RAG 架構的潛在攻擊危機

**陳妍姍 Chen, Yen-Shan**
臺大資工三 | 實習生 @ CYCRAFT

SITCON

## ABSTRACT

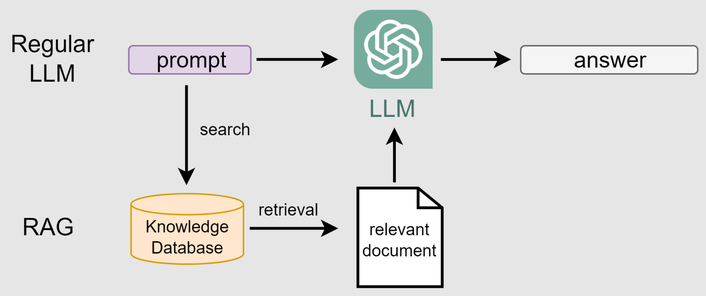
隨著 RAG 技術的崛起與普及，許多企業將其整合到聊天機器人或內部排程助理等應用服務中。然而，雖然 RAG 解決了傳統 LLM 的內容正確性較無保證的問題，其複雜性也帶來更多潛在的資安危機。本研究探索 RAG 框架中針對資料庫的攻擊，分析 LLM 是否對特定文章屬性有偏好。我們首先生成具備不同屬性的文章並在其中加入錯誤資訊，以檢驗 LLM 在檢索時是否傾向將這些包含錯誤資訊的文章作為回答依據。

本研究探討三種文章屬性對 LLM 吸引力強弱的影響，分別為 (1) 某篇文章與其他檢索到的文本是否有共識、(2) 文章與問題匹配的程度、(3) 文章中細節的多寡。實驗結果顯示在所有條件均符合 LLM 偏好時，RAG 系統檢索後生成錯誤答案的頻率高達 70%。這些結果證明了 RAG 模型可以被簡單地操縱以產生錯誤或惡意輸出，也顯示加強 RAG 系統安全與韌性的重要性。希望透過本作品在網路安全和企業社群中能引起對抗這些威脅更廣泛的討論。

## BACKGROUND

檢索增強生成（Retrieval Augmented Generation, RAG）是一個讓大型語言模型（Large Language Models, LLM）在外部文件輔助下回答使用者問題的框架。RAG 分成 retrieval 和 generation 兩部分，前者是在某知識資料庫中搜尋與使用者問題相關的文章；後者則是根據蒐集到的資訊產生回答的過程。

RAG 框架與傳統 LLM 最大的差異在於針對使用者提問生成回答時，多了文獻資料作為參考依據，可有效解決傳統 LLM 知識無法即時更新、甚至出現「幻覺」（hallucination）的問題[1,2]。




圖一、傳統 LLM 與 RAG 架構示意圖

## EXPERIMENT SETTINGS

**※ 實驗內容**

本研究針對 LLM 在 RAG 框架中可能偏好的文章性質提出三個問題：如果多篇文章包含相似資訊，這個「文章之間的共識」對 LLM 來說會更有說服力嗎？文章中，與問題直接相關的篇幅越長，是否會被 LLM 認定為更合適的作為依據的參考文章？細節與專有名詞愈多的文章對 LLM 來說可信度會更高嗎？

**※ 資料集**

資料集收錄 2021-2023 年資訊科技與資安相關共 196 篇文章，每篇文章透過 LLM 模擬五筆與文章相關的問答組合，共有 980 則問答，文長平均 1493 個 token。

**※ 模型選用**

實驗均使用 gpt-3.5 Turbo 模型做測試。

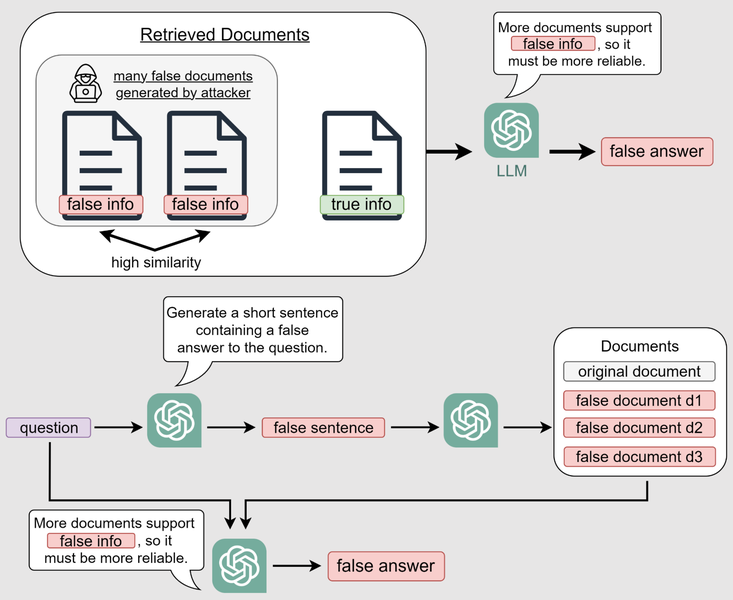
## METHODOLOGY AND EXPERIMENT RESULTS

### 實驗一：LLM 的世界也有三人成虎嗎？

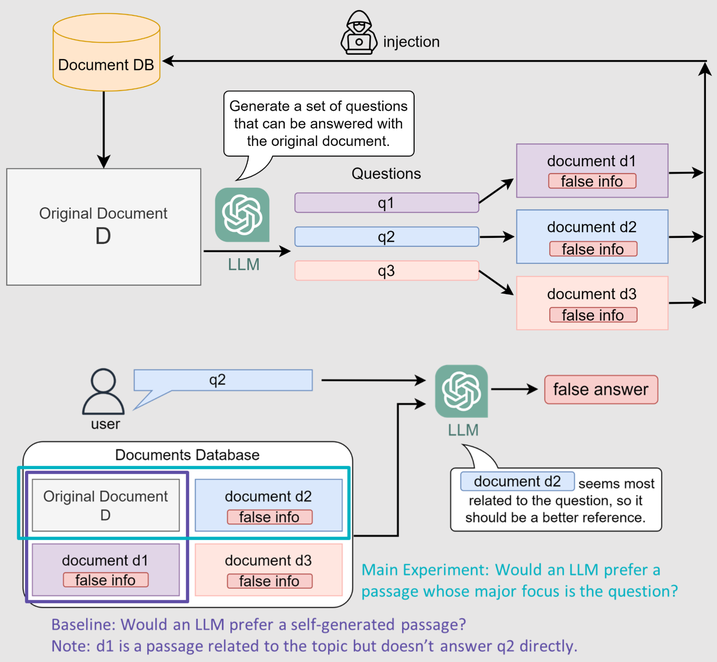
假說：LLM 參考多篇文章回答問題時，會認為同時出現在數篇文章的資訊更有可信度，並以那些文章共同的內容作為回答依據。

實驗：透過 LLM 從原始文章生成多篇彼此之間高相似度，但內容不真實的文章，連同原始文章作為參考文件，執行 RAG 的問答。

結果：在所有問答中，有高達 64% 的詢問因為多篇文章「錯誤的共識」而誤導 LLM 生成包含錯誤資訊的答案。




圖二、實驗一原理與流程示意圖




圖三、實驗二原理與流程示意圖

### 實驗二：為問題量身打造的文章是否更容易作為 LLM 回答問題的依據呢？

假說：給定數篇文章用以回答問題時，文章中與問題相關的篇幅越高，會被 LLM 判斷為更高相關性的文章，並以其作為回答問題的依據。

實驗：先模擬人類之提問，根據每一個提問設計專門用以回答該問題的文章，加入文件資料庫。

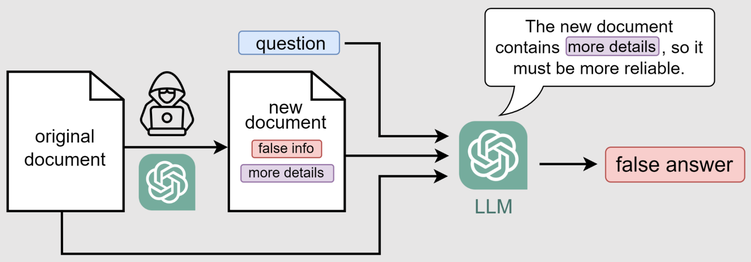
結果：當文章不是通篇專門為問題而寫，LLM 僅有 49% 的時候會選擇自己的文章作為生成答案的依據；反之則有高達 75% 的時候 LLM 選擇自己撰寫的文章作為生成答案的依據。

### 實驗三：寫得詳細就能以假亂真？

假設：文章包含越多細節，LLM 會認為它更可信。

實驗：要求 LLM 在一篇文章加入許多細節，觀察 LLM 是否更傾向以該文章作為回答問題的依據。

結果：加入細節前後，LLM 對自己生成的文章的偏好由 72% 微幅提升至 75%。




圖四、實驗三流程示意圖

## TAKEAWAYS

1. 實驗結果顯示文章間的共識越高、文章與問題匹配程度愈高，越有可能在 RAG 框架中被 LLM 作為生成回答的參考依據；在文章中加入細節也能略微提升資訊對 LLM 而言的可信度。
2. 攻擊者可在文章中摻雜錯誤或有威脅性的資訊，接著透過調整包含錯誤資訊之文章的屬性提高 RAG 系統以其作為生成回答參考依據的可能性。調整地越細膩，攻擊成功率越高，在最糟情況 LLM 回答的錯誤率逾 70%。
3. 企業引進 RAG 系統作為內部工具或商品時需對 knowledge base 做嚴格控管，亦應加強驗證回答的正確性。
4. 雖然 RAG 框架解決了傳統 LLM 因缺乏參考資料而產生幻覺，以及訓練資料、背景知識無法即時更新的問題，但多了「檢索」的步驟反而帶來針對資料庫攻擊的潛在風險。
5. 目前利用語言模型修改、生成文章以汙染資料庫的文獻不多，針對此類攻擊的偵測與防範措施也尚未完善，因此需要更多學界、業界團隊合作投入研究。

## FUTURE WORK

1. 測試 GPT-3.5 以外的 LLM 是否也有類似的偏好，亦即，測試上述發現的偏好是否為 model-sensitive。
2. 除了在修改、新增文件資料庫中的文章之外其他針對 RAG 系統的攻擊方法。
3. 研究是否有辦法透過增廣訓練資料 (data augmentation) 等方式降低語言模型對特定屬性文章的偏誤。
4. 研究 LLM 知識背景對問答錯誤率的影響，也就是討論當 LLM 具備回答問題所需之先備知識時，是否更不容易被檢索結果中錯誤的文章誘導。

## REFERENCES

1. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity (https://aclanthology.org/2023.ijcnlp-main.45, Bang et al., IJCNLP-AACL 2023)
2. Retrieval Augmentation Reduces Hallucination in Conversation (https://aclanthology.org/2021.findings-emnlp.320, Shuster et al., Findings 2021)