國立臺灣大學管理學院資訊管理學研究所

碩士論文

Department of Information Management

College of Management

National Taiwan University

Master's Thesis

檢索增強生成適應性指標研究

RAG Adaptability Metric Study

魏冠宇

Kuan-Yu Wei

指導教授: 曹承礎 博士

Advisor: Seng-Cho Chou, Ph.D.

中華民國 113 年 6 月

June, 2024

doi:10.6342/NTU202401285

摘要

大型語言模型（LLM, large language model）持續快速發展，為彌補模型的不足與因應不同情境下的使用需求，檢索增強生成（RAG, Retrieval-Augmented Generation）等模型已被成熟運用。然而，大型語言模型如何處理「檢索文檔」（Retrieved Documents）尚屬於一個「黑盒子」，其決策過程封閉且不透明，限制了解釋性與可追蹤性。

本論文提出了檢索增強生成適應性指標（RAG Adaptability Metric），藉由提示工程（Prompt Engineering）使生成模型（Generation Model）在生成回答時同時輸出支持文檔（Supporting Documents），讓模型同時指出其認為的產生回答的依據。

本論文提出了檢索增強生成適應性指標（RAG Adaptability Metric），通過提示工程（Prompt Engineering）使生成模型（Generation Model）在生成回答時同時輸出支持文檔（Supporting Documents），讓模型指出其認為的產生回答的依據。

研究中也發現，生成模型在發現檢索文檔不足以支持生成回應時，會轉而採用生成模型訓練過程學習的知識，即記憶化參數來生成回應內容，在部分情境中，記憶化參數生成的回應是可接受的，但同時存在生成幻覺（Generative Hallucination）的風險，因此，本研究藉由取得支持文檔與提問、檢索文檔、回答間的內容相關性產生檢索增強生成適應性指標，賦予檢索增強生成模型，生成過

doi:10.6342/NTU202401285

程的解釋性與可追蹤性。

研究結果顯示，檢索增強生成適應性指標適用於多種檢索方法與不同的生成模型，在識別潛在有風險的生成結果上表現良好，並且可以協助辨別大型語言模型是否依照檢索文檔生成回答，還是發生「拒絕回答」或「自行產生回答」的情境，提供調適模型或訓練模型的參考依據。

關鍵字：檢索增強生成、檢索增強生成適應性指標、檢索增強生成提示工程

doi:10.6342/NTU202401285

Abstract

Large Language Models (LLMs) are continuously advancing at a rapid pace. To address the shortcomings of these models and cater to various contextual usage needs, models like Retrieval-Augmented Generation (RAG) have been maturely utilized. However, how LLMs handle ”retrieved documents” remains a ”black box”, with a decision-making process that is closed and opaque, limiting explainability and traceability.

This paper proposes the RAG Adaptability Metric, which uses prompt engineering to enable the generation model to output supporting documents when generating responses, thus allowing the model to indicate the basis for its responses. The study found that when the retrieval documents are insufficient to support the generated response, the model tends to rely on the knowledge learned during its training process, i.e., memorized parameters, to generate the response content. In some scenarios, responses generated from memorized parameters are acceptable, but there is a risk of generative hallucination.

Therefore, this study introduces the RAG Adaptability Metric by obtaining the rel-

doi:10.6342/NTU202401285

evance between supporting documents, queries, retrieved documents, and responses, enhancing the explainability and traceability of the RAG process. The results show that the RAG Adaptability Metric is applicable to various retrieval methods and different generation models, performing well in identifying potentially risky generated responses. It can help distinguish whether the LLM generates responses based on retrieved documents or if it occurs in situations of ”refusal to respond” or ”self-generated responses”, providing reference for adjusting or training the models.

Keywords: RAG, RAG Adaptability Metric, RAG Prompt Engineering

doi:10.6342/NTU202401285

目次

Page

摘要 **i**

Abstract **iii**

目次 **v**

第一章 **Introduction** **1**

1.1 研究背景 1

1.2 研究動機 3

1.3 研究預期產出 5

1.4 研究假設 6

第二章 **Related Work** **8**

2.1 Retrieval Augmented Generation 8

2.2 Self-RAG 9

2.3 Model Hallucination 9

2.4 Prompt Engineering 10

2.5 Refusal to Answer and L2R 11

2.6 Embeddings 12

2.7 Retrieval Augmented Generation Assessment 12

第三章 **Methodology** **14**

3.1 模型設計 14

3.1.1 相關性閥值（Relevance Threshold） 15

doi:10.6342/NTU202401285

3.1.2 檢索增強生成適應性指標 15
3.1.3 主要指標（Primary Metric） 16
3.1.3.1 （1）支持文檔與提問的相關性指標 16
3.1.3.2 （2）支持文檔與檢索文檔的相關性指標 16
3.1.3.3 （3）支持文檔與回答的相關性指標 17
3.1.4 輔助判斷指標（Auxiliary Assessment Metric） 17
3.1.4.1 （4）提問與檢索文檔相關性 17
3.1.4.2 （5）提問與回答相關性 18
3.1.5 預期實驗設計 19
3.2 實驗模型設計 20
3.2.1 指標數值制定 21
3.2.2 結果評估 21
3.2.3 指標性能實驗 22
3.2.4 生成器實驗 22
3.3 實驗資料集 22
3.3.1 Cloud Platform FAQ Dataset 23
3.3.2 Stanford Question Answering Dataset (SQuAD) 23
第四章 **Results** **24**
4.1 檢索增強生成適應性主要指標性能 24
4.1.1 ROC Curve 24
4.1.2 Analysis of F1-Score and F2-Score 25
4.1.3 其他生成模型 29
4.2 輔助判斷指標性能 31
4.3 區分回答依據 32
第五章 **Discussion** **34**
5.1 實驗結果與應用 34

doi:10.6342/NTU202401285

5.2 大型語言模型性能影響 35
第六章 **Conclusion** **36**
6.1 Limitation 37
6.1.1 相關性分數與閥值的限制 37
6.1.2 僅適用於大型語言模型生成器 38
6.1.3 實驗過程限制 38
6.2 Future work 39
6.2.1 探討提示詞的影響 39
6.2.2 與其他檢索增強生成架構互動的能力 40
6.2.3 相關性閥值與指標權重的調整 40
參考文獻 **42**

doi:10.6342/NTU202401285

第一章　Introduction

1.1　研究背景

在當前生成式 AI 模型（Generative Artificial Intelligence）快速發展的背景下，大型語言模型（LLM, large language model）得到了廣泛的應用。企業和組織在建立基於大型語言模型的應用服務時，通常採用預訓練模型（Pre-trained Model）。預訓練模型在 AI 技術爆發的時代快速推陳出新，其訓練過程中使用了大量的數據，並採用了眾多權威知識來源，如維基百科和政府智庫。因此，大型語言模型在理論上應該具備高度的準確性，但在實際應用中，模型幻覺（Model Hallucination）依然困擾著 LLM 的使用者，並對生成的回答質量有顯著影響。

模型幻覺指的是模型產生隨機且不可預測的回答，或是看似合理但實際上不正確的回答。正如 Ziwei Ji 等人（2022）[5] 所指出的，模型在訓練過程中受到訓練反饋、強化學習（Reinforcement learning）和語義接地（Semantic Grounding）等技術的影響，為了確保生成內容的多樣性和創意性，幻覺的產生在目前生成模型中是必然會發生的。

在實務運用中，企業或組織經常將大型語言模型應用於知識性問答或客戶服務，這些應用不可避免地受到幻覺的影響。此外，預訓練數據的限制也是一個重要問題。例如，「GPT-4」的訓練數據截至 2021 年 9 月，對於這之後發生的事件、

doi:10.6342/NTU202401285

新知識，以及企業或特定專業領域的資料，由於不在預訓練數據集中，模型無法得知。這些都是使用預訓練大型語言模型作為應用基礎時，不可避免的限制和問題。

因此，實務上，大多數大型語言模型應用服務採用「檢索增強生成模型」（Retrieval-Augmented Generation, RAG）（Patrick Lewis, 2020）。該技術透過檢索外部資訊作為大型語言模型的輸入，減少幻覺的產生，提高回答的品質，並且能夠提供大型語言模型未學習的知識。常見應用包括企業客戶服務機器人或特定領域知識問答。

由 Patrick Lewis 等人（2020）[7] 提出的檢索增強生成模型由兩個部分構成：檢索器（Retriever）與生成器（Generator），大部分檢索增強生成模型的生成器採用大型語言模型，在模型運作過程中又稱為生成模型，而檢索器則會利用餘弦相似度（Cosine Similarity）等搜尋相關文本內容（Context）的方法，提取適用於協助回答問題的檢索文檔（Retrieved Documents）。在取得檢索文檔後，通過嵌入合併（Embedding Fusion），或是提示工程（Prompt Engineering）將檢索文檔（Retrieved Documents）與提問內容（Question）做為大型語言模型的輸入，以取得準確性較高的回答。

即使檢索增強生成技術對於大型語言模型應用服務提供顯著的改善，但如 Jiawei Chen 等人（2023）[2] 在其研究中所評估，檢索增強生成技術還是有許多改進空間，例如，雖然生成模型有機率對於品質不良的檢索文檔進行「拒絕回答」（Refusal to Answer），但其表現仍受大型語言模型性能的影響。同時有機率發生，即使文檔中明顯沒有可以參考的內容，模型依然會依據文檔產生似是而非的回答，正如 Minhyeok Lee（2023）[6] 指出的那樣，在沒有明確答案的情況下，生成模型難以做出判斷，提示詞提供的上下文（Context）也會影響生成字彙的權

doi:10.6342/NTU202401285

重，因此在檢索文檔（Retrieved Documents）品質不良，包含過多噪音（Noisy）、與提問不相符的內容或無法支持回答問題時，反而會使得回答品質變得更差。

之所以會出現這種現象，也在 Theodore Zhao 等人（2023）[17] 的研究中提出，大型語言模型在訓練過程中，由於強化學習的因素，傾向於給出正面的回答（Positive Response）而非拒絕回答，這是因為大型語言模型尚未具備足夠的認知能力來審視知識的正確性。在大型語言模型內部進行生成決策時，為了盡可能提供正面回答，模型有可能會採用部分檢索文檔的內容以及訓練過程中所學習的知識和記憶化參數來生成回答。在這種情況下，很難確認回答的準確性以及模型幻覺是否發生，因此，當檢索文檔的品質不佳或無關時，生成的回答可能會更不可信賴，進而影響整體回答品質。

雖然有這些問題，但 Yunfan Gao 等人（2023）[4] 也在論文中表明，檢索增強生成技術在近期仍蓬勃發展，因為其高度彈性與適應性，儘管這些模型各自存在限制並受到幻覺的困擾，但檢索增強生成技術的穩健性和處理延伸上下文的能力，使其在實務上能帶來很大的協助並受到業界的廣泛應用，目前仍有許多研究在繼續推動檢索增強生成技術的進步。

1.2 研究動機

Jiawei Chen（2023）[2] 在實驗過程中發現，即使檢索文檔充分支持回答問題，但生成模型仍有可能採用記憶化參數作為回答。原始的檢索增強生成模型僅能取得提問、檢索文檔、回應三個部分，這其中的關聯性並沒有一個可靠的邏輯，且在生成過程中，因為預訓練大型語言模型「黑盒子」的特性，使用者無法得知其中的判斷依據，並且缺乏可解釋性和可追蹤性，無法確定模型是使用檢索文檔的內容還是記憶化參數來回答問題，也無法在生成結果出現問題時，快速釐清問

doi:10.6342/NTU202401285

題發生的原因。

企業或組織在運用檢索增強生成模型時也常遇到此問題，且缺乏一個標準來辨別模型的回答是否根據檢索文檔，以及模型幻覺是否發生。Akari Asai 等人（2020）[1] 提出了 Self-RAG 模型，這也是部分企業或組織對大型語言模型應用所採用的解決方法，雖然可以有效改善模型輸出的回應品質，但仍無法完全遏止模型幻覺的產生。並且，由於生成過程中多次進行檢索和反思，計算資源需求或時間複雜度較高，可能會導致回應速度下降和運行成本上升。在需求變化或大型語言模型更新時，需要重新調適模型或調整結構。

Sina J. Semnani 等人（2023）[12] 在論文中提到，利用多次檢索和「Few-Shot Grounding」等技術，設計不同的提示工程逐步生成並反覆確認生成內容，可以提高語言模型輸出結果的準確性。然而，由於其模型複雜性及大量使用 Token 所衍生的成本，以及在不同使用情境下需要進行多個面向的調整，不一定符合目前業界或組織設計大型語言模型應用服務時的限制與考量。

因為企業或組織導入檢索增強生成模型時具有風險，錯誤的回答或模型幻覺發生時對檢索增強生成的應用服務影響很大，除了影響使用者的實際體驗與應用效率，做為對外服務甚至可能對商譽造成不佳影響。目前有許多研究開發的檢索增強生成的衍生模型，大多具有較高的時間複雜度與成本消耗，並且需要投入較多額外開發資源。在不確定檢索增強生成模型帶來的效益前，企業難以選擇適合的檢索增強生成結構進行導入，並做出投入資金、時間與開發成本的決策，特別是對於資源有限的組織而言。

另外，因為大型語言模型生成的內容難以有一個客觀的衡量標準，所以在訓練模型或微調模型時很難有協助訓練的標準，如果有一個指標能協助評量檢索增強生成過程的品質並給予客觀分數，則可以用於調適模型生成品質。

doi:10.6342/NTU202401285

1.3 研究預期產出

本研究希望能夠證明「藉由提示詞來探討檢索增強生成模型中，生成模型對於檢索文檔的使用狀況」是有意義的，並提供一個適用於檢索增強生成模型的適應性指標（RAG Adaptability Metric）。

檢索增強適應性指標是衡量模型在檢索以及生成方面的適應性，評估檢索過程資料的上下文與語意一致性，以及各階段內容的相關性，生成回答的精確性以及錯誤發生時的偵錯能力等。

在相較於其他檢索增強生成延伸模型較低的時間複雜度與 Token 消耗的情況下，提供相對精準的判斷依據，讓模型開發者可以快速調適模型或判斷模型輸出內容品質，不會因為檢索方式的調整或生成器採用的預訓練大型語言模型更換而需要重新訓練。

檢索增強生成適應性指標是開發友善的，不需要使用大量資源與時間調適模型，即可提供判斷依據，以及賦予檢索增強生成可追蹤性與可解釋性，藉由指標內容判斷生成器是否依據檢索文檔進行回應生成，或由大型語言模型中記憶化參數生成，以及排除生成過程品質不佳的生成結果，或賦予良好的生成結果可讀性。

檢索增強生成適應性指標是藉由修改提示詞，讓大型語言模型在輸出回答時，同時輸出大型語言模型認為用來產生回答依據的支持文檔，藉由探討提問、檢索文檔、支持文檔、回答之間的相關性，賦予檢索增強生成過程可追蹤性、可解釋性的指標，研究中會探究這些指標的效能與意義，並驗證藉由提示詞探究大型語言模型產生回答的決策過程是否可行。

doi:10.6342/NTU202401285

1.4 研究假設

本研究認為，藉由提示工程提取得支持文檔（Supporting Documents），對於了解生成模型的認知，以及對檢索文檔的使用狀況是具有價值的。

支持文檔是通過在輸入時加入「Provide the content that you think supports the answer」等提示詞，使生成模型在產生回答時，同時輸出其認為能支持回答的內容。

如 Jason Wei 等人（2022）[13] 所提出「湧現現象」（Emergent Abilities），大型語言模型在訓練量增長時逐漸衍伸和強化某些能力，包括推理及辨別是非的能力。這些能力可以藉由適當的提示工程引導模型回答出推理過程。

Jason Wei 等人（2022）[14] 也在另外一篇論中也提出「思維鏈」（CoT, Chain-of-Thought），指出雖然大型語言模型具有邏輯思維，但如果沒有加上逐步思考的概念，則可能在推理過程中發生錯誤。因此，本研究嘗試將大型語言模型依據檢索文檔生成回應的過程拆解出來，賦予大型語言模型支持文檔（Supporting Document）的概念。讓大型語言模型在回答時同時輸出「它」認為產生回答的依據。

如果回答受到檢索文檔的影響較大，則支持文檔會與檢索文檔有較高的相關性；反之，如果檢索文檔中的內容不足以支持回答，大型語言模型會利用記憶化參數產生支持文檔，或做出「拒絕回答」的回應。

因此，本研究認為探討支持文檔與其他檢索增強生成模型中的內容相關性是有意義的，可以從中評估檢索增強生成過程的品質是否良好，並觀察生成過程中的各項指標，理解檢索增強生成模型在生成過程中的表現，並協助釐清表現不佳的原因。

doi:10.6342/NTU202401285

本研究認為，在檢索增強生成模型中，生成器產生「拒絕回答」或沒有依照檢索文檔而「使用記憶化參數生成回答」時，可以通過檢索增強生成指標觀察到這些情況。當支持文檔與檢索文檔相關性分數較低，但提問與回答同時具有高度相關性分數時，這意味著大型語言模型認為檢索文檔的品質不佳而自行生成支持回答的內容。而當拒絕回答發生時，產生的回答內容會與提問或檢索文檔差異較大，回答與支持文檔、支持文檔與檢索文檔的相關性分數會比較低。

檢索增強生成指標是在修改提示詞的基礎下，用取得的各項資料計算出來的，因此在替換不同的大型語言模型生成器時，應該都可以正常運作，並且使用邏輯能力越強的生成模型應該要有更好的表現。

doi:10.6342/NTU202401285

第二章　Related Work

2.1　Retrieval Augmented Generation

由 Facebook AI Research (FAIR) 中 Patrick Lewis 等人（2020）提出的自然語言處理架構，將自然語言生成過程分為檢索器（Retriever）以及生成器（Generator），檢索器主要用於取得查詢（Query）中所需的相關信息，讓生成器具備更完整的知識生成回答。

也驗證檢索增強生成可以有效改善模型幻覺（Model Hallucination），藉由檢索器取得之檢索文檔（Retrieved Documents），改善自然語言模型中預訓練知識更新較慢的問題，以及減少模型採用訓練過程學習的記憶化參數作為回應，研究結果展示使用者更傾向於使用檢索增強生成模型產生的回答，且優於僅使用模型記憶化知識作為生成回答的模型。

即便檢索增強生成在處理自然語言任務上取得很大的進步，但在特定資料集的問答上還是半數以上的機率表現不佳。

後續 Akari Asai 等人 (2023)[1] 在研究中指出單獨使用檢索增強生成會降低大型語言模型的多功能性，並限制生成內容，因為檢索增強生成會大規模的檢索文檔，不論這些文檔對於產生回應是否有幫助，或者是檢索出與主題無關或不必要的文檔，從而影響大型語言模型產生回應的 Token 權重，導致生成質量降低，此

doi:10.6342/NTU202401285

外，檢索增強生成的回應輸出內容依據不一定與檢索到的文檔內容一致，因為大型語言模型並沒有被特別訓練於遵循或使用這些輸入的檢索文檔。

2.2 Self-RAG

Akari Asai 等人（2023）[1] 提出了「Self-RAG」模型，是一個檢索增強生成自性應、自我改進的架構，他們在研究中額外訓練一個模型，根據任務輸入生成反思標記（Retrieval 和 Critique），並利用反思標記來判斷是否需要檢索或評估生成質量，此外，Self-RAG 會並行處理多個檢索文檔，評估相關性後將有用的文檔片段作為大型語言模型的輸入，同時也會排除無意義或錯誤的文檔片段，以確保回答的正確性以及確保整體質量。

Self-RAG 要求大型語言模型在生成過程中進行自我反思和批評，這增加了系統的複雜性。如果不另外訓練反思模型，則依然會遇到模型幻覺的問題。此外，反思標記的表現直接影響生成結果的品質，如果反思標記的準確性不佳，可能導致檢索過程中的決策錯誤。因此，反思標記的調適和訓練可能需要花費大量的時間和人力，並且需要在需求變化時進行調整。

2.3 Model Hallucination

Minhyeok Lee(2023)[6] 在關於模型幻覺的研究中提到，大型語言模型，特別是以自監督式學習（Self-Supervised Learning）為基礎，會利用大量的未標記數據進行訓練，通過預測序列中的後續標記來生成和處理自然語言，其目標是在提升給定模型參數的情況下產生數據的可能性。

Ziwei Ji 等人 (2022)[5] 在研究中指出，模型幻覺是大型生成式模型（Genera-

doi:10.6342/NTU202401285

tive Model）的共通特性，正面用途在於提供變化性與差異性，並能產生不重複的內容。

Minhyeok Lee(2023) 也提到幻覺與創造力之間具有錯綜複雜的關係，是大型語言模型固有限制的結果，藉由修改模型來完全遏制模型幻覺的產生是困難且缺乏效率的，但在論文中也指出了幾個幻覺發生的主因，包含語意不清的上下文（Context）、複雜的提示以及訓練過程中權重較低或未知的知識內容。

Jacob Menick（2022）等人 [8] 在先前的研究中也提到，儘管語言模型時常可以正確回答問題，但用戶不能相信語言模型的任何主張，模型幻覺時產生的回應時常是似是而非的答案，儘管看似不合理或與真實世界事實不符，但大型語言模型仍會視為合理輸出作為回應，反而貼近真實情況的回應會因為大型語言模型產生回應時，預測權重較低而未被輸出。

2.4 Prompt Engineering

Orion Weller 等人（2023）[15] 提到大型語言模型存在幻覺並生成虛假信息的問題，即使預訓練的資料是真實存在並且正確的，甚至具有一定的參考依據，大型語言模型仍可能會杜撰出它認為更好的答案，或在正確的回答中參雜部分虛假的信息。該論文中提出了一個改善方法，利用「根據來源」（According to Sources）的概念，通過提示工程改變語言模型預測生成內容的權重。例如，在提示（Prompt）中添加「根據維基百科（Wikipedia）」，大型語言模型會更偏好採用預訓練中維基百科的資料生成回答內容。

為了量化並衡量此提示工程的效果，Orion Weller 等人提出了一種新的評估指標，QUIP-Score（Quoted Information Precision Score），該指標衡量大型語言模型生成的答案在預訓練文檔資料集中能直接找到的程度。

doi:10.6342/NTU202401285

通過維基百科問答集的實驗表明，「根據來源」可以提升大型語言模型在限制來源情況下的表現，使 QUIP-Score 的分數上升。他們也通過提示大型語言模型減少限制，或根據其他資料來源來產生回答，觀察到回答的 QUIP-Score 下降，從而驗證大型語言模型確實會因為提示詞的不同而改變生成內容對於文檔來源的依據。

Sarah Wiegreffe 等人（2021）[16] 在研究中表示，可以通過提示詞讓大型語言模型生成解釋這些回答的文本，這些解釋有時是有意義的，但在正確性以及因果關係上仍有改進的空間，並且無法驗證解釋內容的真實性。

2.5 Refusal to Answer and L2R

「拒絕回答」（Refusal to Answer）是大型語言模型知道無法可靠的生成回答，而表示無法或拒絕作答的行為。Akari Asai 等人 (2023)[1] 在論文中也指出，檢索增強生成模型中生成模型很少因為支持文檔依據不足，而做出「拒絕回答」，Sina J. Semnani 等人（2023）在研究中也是使用「Grounding」以及提示詞讓大型語言模型做出拒絕回答的回應，以及 Theodore Zhao 等人（2023）在論文中提及大型語言模型的自我校正，但依然無法保證大型語言模型做出拒絕回答的穩定性。

Lang Cao（2023）提出一種 L2R（Learn-to-Refuse）的方法，結合結構化知識庫和自動知識擴展的方法，設立的大型語言模型知識範圍的界線，讓大型語言模型具在知識範圍之外可以做出拒絕回答的決策，論文中也提及利用 L2R 訓練一個判斷是否可以回答的機制，如果可以回答的話需要包含證據（Evidence）和原因（Reasoning），這個機制使問答系統更具有可靠性。但在論文中也提及，L2R 仍無法遏止幻覺的產生，並且需要花費較多的時間與開發資源進行知識庫的建立，距離直接做為應用還有一段努力要做。

doi:10.6342/NTU202401285

2.6 Embeddings

詞嵌入在語義檢索、文本相似度計算和代碼搜索等多種應用中是判別特徵的關鍵因素。先前的詞嵌入通常根據不同的使用情境定義模型，其中包含資料集的選用、訓練目的以及模型架構等差異。

Arvind Neelakantan 等人（2022）[9] 提出對比預訓練（Contrastive Pre-Training）的方法，使用無監督式學習的方式進行大規模的訓練，該方法使用 Transformer 編碼器對輸入文本進行編碼，並在沒有額外標籤的情況下，利用自然配對的資料構建訓練集。

經過對比預訓練文本嵌入模型在語義搜索和文本相似度計算中表現出色，並且在各項分類任務中表現出出色的結果。此外，在大規模語義搜索任務中，該模型在 MSMARCO、Natural Questions 和 TriviaQA 的表現相對於之前的方法也都有顯著提升。

對比預訓練中也能將長短不同但語意相近的句子映射到相似的向量空間，將對比學習目標（Contrastive Learning Objective）中語義相似的句子最大化，以及最小化與其他句子的相似性來訓練模型，以保持對長度差異較大的字句間，語義的準確理解。

2.7 Retrieval Augmented Generation Assessment

RAGAS（Retrieval Augmented Generation Assessment）是由 Shahul Es 等人（2023）[3] 提出的一種評估方法，通過評估信實性（Faithfulness）、答案相關性（Answer Relevance）與上下文相關性（Context Relevance），來衡量檢索增強生

doi:10.6342/NTU202401285

成模型的回答內容的可靠度。

信實性利用額外的大型語言模型請求來評估答案內容是否可以從檢索文檔中推導出來。答案相關性則關注答案是否清楚明白地回答了提問。上下文相關性是從檢索文檔中提取關鍵字句，檢查回答是否包含這些關鍵字句。

RAGAS 的研究表明，探究提問、檢索文檔、回答三者間的相關性是有意義的。然而，RAGAS 的方法在評估上非常依賴用以協助判斷的大型語言模型性能，並且在評估過程中需要耗費大量計算資源與成本，因此在實務運用上還有很大的提升空間。

doi:10.6342/NTU202401285

第三章　Methodology

3.1　模型設計

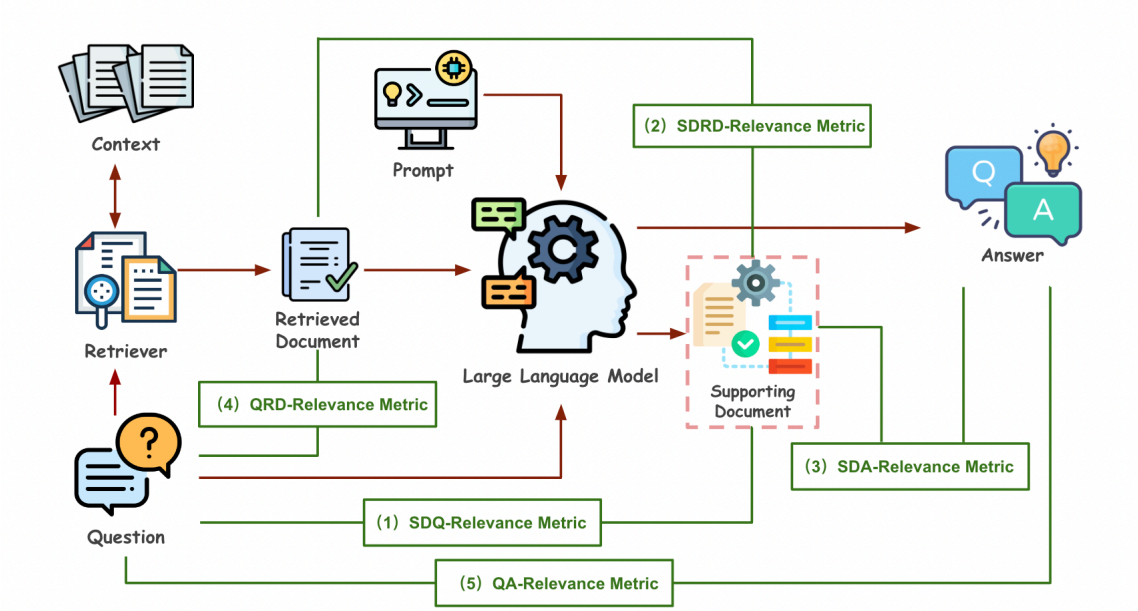

Figure 3.1: 檢索增強生成模型與適應性指標，紅色為檢索增強生成模型執行的流程，綠色部分為適應性指標及其變量

原有的檢索增強生成模型，通過提問（Question）或 Query 取得檢索文檔（Retrieved Document），並使用大型語言模型做為生成模型，根據檢索文檔以及提示詞（Prompt）產生回應內容。本研究在原有提示詞中添加內容「Provide the content that you think supports the answer」，引導大型語言模型輸出支持文檔（Supporting Document）。

doi:10.6342/NTU202401285

3.1.1 相關性閥值（Relevance Threshold）

相關性閥值是用來定義指標表現為正類（Positive）或負類（Negative）的標準，若指標相關性分數大於相關性閥值，則表示此指標代表的內容在該次任務中表現良好，低於標準值為表現不佳。

3.1.2 檢索增強生成適應性指標

檢索增強適應性指標，是將檢索增強生成模型生成過程間的文本內容，進行相關性的比對，指標代表兩個變量間的相關程度，並用指標數值是否大於相關性閥值來定義表現為良好還是不佳。定義支持文檔與提問、檢索文檔、回答之間的相關性以及提問與檢索文檔、回答之間的相關性為檢索增強適應性指標的內容，如 Figure1.1 綠色部分所示。

根據 Shahul Es 等人（2023）[3] 的研究中指出，檢索增強生成模型過程間的文本內容具相關性有因果性與可解釋性，因此適應性指標可以反映檢索增強生成模型在不同情境下的適應能力，幫助理解模型在特定情境中的表現，同時也能指出模型在哪些情境下可能產生誤差或幻覺。

指標數值的計算方式採用 Arvind Neelakantan 等人（2022）所提出的衍生模型「text-embedding-3」作為詞嵌入模型，經過詞嵌入處理轉成向量後，使用餘弦相似度（Cosine Similarity）計算變量間的相關性分數，並通過相關性分數與相關性閥值計算該指標為正類或負類。

指標依照泛用性（Generality）及功能性區分為主要指標（Primary Metric）與輔助判斷指標（Auxiliary Assessment Metric）。主要指標是支持文檔與檢索生成增強模型中的上下文內容相關性；輔助判斷指標則是提問與檢索文檔的相關性，以

doi:10.6342/NTU202401285

及提問與回答的相關性。

3.1.3 主要指標（Primary Metric）

主要指標旨在使用生成模型認知的支持文檔來賦予檢索增強生成的回答以可解釋性與透明性，並釐清支持文檔是否與回答相關，而非大型語言模型生成的隨機內容。主要指標強調泛用性以及準確性，用以探討生成模型對於檢索增強生成過程的理解。

3.1.3.1 （1）支持文檔與提問的相關性指標

支持文檔與提問的相關性指標（SDQ-Relevance Metric, Supporting Document-Question Relevance Metric），評估生成模型認知的支持文檔與提問是否具有相關性，用以檢驗生成模型是否理解提問內容或提問意圖，選用適當的支持文檔。

如 Figure 1.2 中示例第二筆，SDQ-Relevance Metric 為 0.77，代表生成模型認定的支持文檔品質不佳，不適合用來回應提問。在考量其他指標的因素下，可以總結出檢索文檔品質不佳，但生成模型依然採用作為支持文檔。

3.1.3.2 （2）支持文檔與檢索文檔的相關性指標

支持文檔與檢索文檔的相關性指標（SDRD-Relevance Metric, Supporting Document-Retrieved Document Relevance Metric）評估檢索文檔在輸入生成模型後，是否被生成模型當作產生回答的依據或認定可以支持回答。

如 Figure 1.2 中示例第一筆，生成模型如果認為檢索文檔適合作為支持回答的依據，則輸出檢索文檔的部分內容作為支持文檔。

doi:10.6342/NTU202401285

如果大型語言模型自行修改檢索文檔內容，或使用記憶化參數自行生成支持文檔，則較低的相關性會反映在指標值中。

3.1.3.3 （3）支持文檔與回答的相關性指標

支持文檔與回答的相關性指標（SDA-Relevance Metric, Supporting Document-Answer Relevance Metric）評估生成模型輸出的回答是否依照支持文檔的內容。由於大型語言模型的特性及幻覺發生的可能性，模型可能會自行產生與支持文檔無關的回答內容。

如 Figure 1.2 中示例第二筆。如果大型語言模型做出「自行生成回答內容」，也可以藉由相關性分數顯著低於標準值來辨別因為生成模型雖然採用了支持文檔，但最後回答並沒有依照支持文檔，且提問與回答的相關性明顯高於提問與支持文檔相關性。

如 Figure 1.2 中示例第五筆，支持文檔參考了檢索文檔，但生成器認為支持文檔無法作為回答的依據，因此回答的內容明顯與支持文檔內容不同。

3.1.4 輔助判斷指標（Auxiliary Assessment Metric）

輔助判斷指標用於評估檢索文檔的品質是否良好，以及提問與回答是否具有基本的相關性。輔助判斷指標缺乏泛用性，並具有一定程度的不穩定性，但依然可以協助判斷特定任務的表現。

3.1.4.1 （4）提問與檢索文檔相關性

提問與檢索文檔相關性指標（QRD-Relevance Metric, Question-Retrieved Document Relevance Metric）用於評估檢索文檔的品質，或檢索過程是否存在異常。該

doi:10.6342/NTU202401285

指標需要根據檢索方式設計來制定評估標準，在基於提問字詞相關性來取得檢索文檔的模型中，該指標效果良好；然而，若檢索器的設計並非基於語意或字詞內容相關性來檢索文檔，則該指標不適用。例如，通過使用者編號或信箱來取得使用者相關資訊的檢索方法，該檢索文檔與提問內容的相關性反而會造成錯誤判斷，因為其泛用信較低，所以將該指標定義為輔助判斷指標。

3.1.4.2 （5）提問與回答相關性

提問與回答相關性指標（QA-Relevance Metric, Question-Answer Relevance Metric）用於檢視回答的內容是否與提問保持一定程度的相關性。高相關性不能保證回答的品質優良，有可能是大型語言模型產生的模型幻覺，輸出似是而非的內容，或未依據模型應用設計者的期待採用檢索文檔。

另一種情況是大型語言模型簡化回答內容，例如提問為「是否可以...」，模型在參考檢索文檔內容後，可能簡短回答「是的，可以」或「不可以」。此情況雖然檢索增強生成模型運作良好，但仍可能因該指標造成錯誤判斷。

然而，此指標仍可用於檢測「拒絕回答」與生成模型「自行生成回答內容」的行為。例如，「QA-Relevance Metric」相關性很高，但主要指標分數較低，則可能是模型採用訓練過程中學習的知識，自行生成回答內容。

雖然此指標對於判斷生成內容準確性的影響較低，但它仍具有協助判斷以及賦予解釋性的價值，因此定義為輔助判斷指標。

Figure 1.2 由上至下展示了適應性指標在不同情境下的表現。

示例第一筆，是理想中的檢索增強生成結果，檢索文檔與提問有高度相關，並被生成模型採用做為支持文檔，回答也依照支持文檔的內容產生。

doi:10.6342/NTU202401285

Question	Retrieved Document	Supporting Document	Answer	(1) SDQ	(2) SDRD	(3) SDA	(4) QRD	(5) QA
Will the nitro hypervisor change the apis used to interact with ec2 instances?	*no, all the public facing apis for interacting with ec2 instances that run using the nitro hypervisor will remain the same*	*In aws service, no, all the public facing apis for interacting with ec2 instances that run using the nitro hypervisor will remain the same.*	*No, the nitro hypervisor will not change the APIs used to interact with EC2 instances.*	0.94	0.99	0.94	0.93	0.93
In ibm service, can i install my own hadoop stack components?	*yes, you can use the hardware development kit hdk either in an aws-provided ami, or in your on-premises development environment*	*you can use the hardware development kit (HDK) either in an AWS-provided AMI or in your on-premises development environment.*	*No, you cannot install your own Hadoop stack components in IBM service.*	0.77	0.98	0.75	0.78	0.92
In ibm service, how do i create a webhook?	*to receive email notifications for a particular topic, a subscriber should specify "email" or "email-json" as the protocol and provide a valid email address as the end-point*	*The retrieved document does not provide information on creating webhooks specifically in the IBM service, therefore the answer is based on general knowledge of webhook creation.*	*you can create a webhook by specifying the appropriate protocol and providing the desired endpoint URL for the webhook.*	0.87	0.76	0.85	0.80	0.93
In ibm service, what regions are bare metal servers for vpc available?	*vcpu-based instance limits are available in all commercial aws regions*	*The retrieved document does not contain information about IBM service or the availability of bare metal servers for VPC in specific regions.*	*No information available in the retrieved document.*	0.86	0.79	0.86	0.84	0.72
In ibm service, does the nlb support layer 7 switching?	*we currently support enhanced networking capabilities using sr-iov (single root i/o virtualization)*	*Enhanced networking capabilities using SR-IOV enable high performance networking for supported instance types.*	*No, the retrieved document does not provide info about NLB supporting layer 7 switching.*	0.78	0.89	0.74	0.79	0.89

Figure 3.2: 雲服務平台 FAQ 資料集中，實際問答內容節錄以及其對應的適應性指標，相關性閥值為 0.8，低於閥值的標示為紅色，高於閥值的標示為綠色

示例第二筆，展示了檢索文檔品質不佳，生成模型雖然選擇檢索文檔做為支持文檔，但在生成回答內容時沒有依照支持文檔，自行生成回答內容。

示例第三筆，展示生成模型認為檢索文檔品質不佳，無法做為回答提問的依據，因此同時表示出檢索文檔的問題，以及自行產生支持文檔與產生回答的情境。

示例第四筆，是生成模型因為檢索文檔品質不佳而產生「表明無法回答的支持文檔」以及「拒絕回答」的行為。

示例第五筆，是生成模型採用部分檢索文檔內容生成支持文檔，但在生成回答時，發現支持文檔無法做為依據，而產生「拒絕回答」行為。

3.1.5 預期實驗設計

在定義好上述指標後，本研究認為，良好的檢索生成增強過程應如 Figure 1.2 的第一筆資料所示，在三個主要指標中都應該具有較高的分數表現，即大型語言

doi:10.6342/NTU202401285

模型能夠理解提問意圖，提供高度相關的支持文檔，檢索文檔被大型語言模型採用作為支持文檔，並且支持文檔被大型語言模型採用作為回答依據。

為驗證這一假設，本研究將採用具有正確答案與檢索文檔的資料集，以及模擬實際檢索增強生成的問答資料集，提供不同的提問以及參考文檔，來檢驗不同情境下檢索增強生成適應性指標的變化。

本研究認為，適應性指標能夠有效反映檢索生成過程中的質量變化，並提供有價值的數據用以改進檢索增強生成模型的生成品質。

除了之外，實驗包含檢測是否能藉由適應性指標分類生成模型「自行生成支持文檔」或產生「拒絕回答」，以及釐清個別指標召回負類樣本的能力。並嘗試替換不同的生成器，觀察檢索增強適應性指標的表現如何。

3.2 實驗模型設計

本研究將建構一個基礎的檢索增強生成模型。首先，對所有參考資料集中文檔片段進行詞嵌入處理（Embedding），計算出參考資料集中最相似的文檔內容，並與提問合併做為生成器的輸入內容。而參考資料集會根據實驗需求進行調整與設計。

同時修改提示詞內容，在生成器輸入時添加「Provide the content that you think supports the answer」這段文字，讓大型語言模型在回答時同時輸出支持文檔，輸出時利用提示工程讓生成器最後輸出內容為「JSON」格式，通過此格式區分支持文檔與回答。

對提問、檢索文檔、回答以及參考文檔使用「text-embedding-3」模型進行詞嵌入處理，並以餘弦相似度（Cosine Similarity）計算參考文檔與提問、檢索文檔、

doi:10.6342/NTU202401285

回答之間的相似度，以及提問與檢索文檔、提問與回答之間的相似度，並將相似度以相關性閾值區分出正負類標記，即檢索增強生成適應性指標。

3.2.1 指標數值制定

為了定義適應性指標的正確或不正確的相似度閾值，本研究選用 SQuAD 資料集。首先，不採用檢索增強生成模型中的檢索器，而是以其文章片段作為檢索文檔，將檢索文檔與關於該片段的提問作為生成器的輸入，取得檢索增強生成適應性指標。

SQuAD 資料集每個提問都有標示是否可以從文章片段中取得回答。如果無法從文章片段中取得回答的提問，其檢索增強生成適應性指標應該能反映出異常。本研究將實驗不同閾值下，對於負類樣本，即無法被回答的問題的「F1-Score」與「F2-Score」，以便根據需求定義適合的閾值。

3.2.2 結果評估

實驗結果將採用「ROC Curve」作為模型整體評估的方法，同時採用「F1-Score」以及「F2-Score」作為評測標準。通過「F1-Score」，我們可以同時考量使用指標判斷正負類樣本的能力，包括精確性（Precision）以及召回率（Recall）。而在研究中增加「F2-Score」的標準，則是因為如 Patricia Craja 等人（2020）在偵測詐欺文檔中的做法，F2-Score 更強調召回率的重要性，因為錯誤的回答可能造成的危害會比不能回答更大。同時監控其他指標也有助於釐清適應性指標的應用情境以及各方面表現。

doi:10.6342/NTU202401285

3.2.3 指標性能實驗

本研究採用雲服務平台 FAQ 資料集，建構一個雲服務平台 FAQ 的檢索增強生成模型，以測試檢索增強生成適應性指標的性能，包括正確率（Accuracy）、辨別檢索文檔缺失的準確率（Precision）、召回率（Recall）、「F1-Score」和「F2-Score」。

實驗過程中僅提供 AWS 的參考資料作為檢索資料，但提問內容涵蓋 AWS、GCP、IBM 三個雲服務平台。如果提問涉及 GCP 或 IBM 平台，則檢索文檔應不適用於協助產生回答。通過這樣的設計，可以驗證檢索增強生成適應性指標是否能協助辨別檢索增強生成模型的檢索資料，是否存在異常或不足以支持回答，並偵測大型語言模型是否自行生成回答內容與支持文檔。

3.2.4 生成器實驗

本實驗將替換不同的大型語言模型作為生成器，觀察 GPT、Gemini、Claude、Llama3 是否適用於檢索增強生成適應性指標，觀察其在相同實驗標準下的表現。這個實驗中採用 SQuAD 資料集，並排除檢索過程的影響。

3.3 實驗資料集

本研究採用兩種不同的資料集，利用雲服務平台 FAQ 資料集限制檢索增強生成模型能取得的檢索文檔內容，檢驗模型對於已知知識但無法取得檢索文檔時的反應，以及檢索增強生成適應性指標是否能反映出這些行為。SQuAD 資料集則可以檢驗，當提供的檢索文檔具有高度相關性但無法作為回答提問的依據時，檢索增強生成適應性指標的反應。

doi:10.6342/NTU202401285

3.3.1 Cloud Platform FAQ Dataset

本研究採用採用 Qdrant（https://qdrant.tech/）提供的雲服務平台 FAQ 資料集「Cloud Platform FAQ Dataset」[10]，Qdrant 是一個開源向量資料庫，主要為 AI 應用服務訓練任務設計。該資料集蒐集了多個雲服務商網站（包括 AWS、GCP、Azure、Hetzner、IBM）的 FAQ，包含服務來源（Source）、匹配的提問（Question）和回答（Answer）。

資料蒐集時間為 2021 年，因此其內容為本研究中大型語言模型預訓練過程中學習的知識，是記憶化參數的一部分。

此資料集可以用於檢測檢索增強生成模型在處理已知知識時的反應，以及檢索增強生成適應性指標對此情境的有效性及準確性。

3.3.2 Stanford Question Answering Dataset (SQuAD)

SQuAD 由 Pranav Rajpurkar 等人 [11] 所提出，是一個用於大型語言模型閱讀理解任務的資料集，該資料集蒐集維基百科中的文章，並進行篩選，擷取文章的片段進行提問。

與其他資料集不同的是，SQuAD 中的提問有很多相似的內容，且部分提問無法從文章中獲得解答，需要通過推理才能獲得答案，因此適用於檢視大型語言模型的閱讀理解能力。

該資料集包含文章片段（Context）、關於該片段的提問（Question），以及該提問是否不可能被回答（isImpossible），即文章片段中無法找到回答的依據。

doi:10.6342/NTU202401285

第四章　Results

4.1　檢索增強生成適應性主要指標性能

以實際雲服務檢索增強生成模型為基礎，採用「雲服務平台 FAQ 資料集」（Cloud Platform FAQ Dataset）做為實驗資料集，選取資料集中「Source」為 AWS 服務的問答集做為參考文檔，並隨機從問答集中抽選 400 題提問（Question）做為樣本，其中 153 筆平台為 AWS（Amazon Web Services）相關問題是可以被回答的，即參考文檔中可以找到與提問有關的對應資料，定義為「正類樣本」。

另外 247 筆為平台為 GCP（Google Cloud Platform）、IBM（IBM Cloud）的相關問題定義為「負類樣本」，因為參考文檔中並不包含這兩個雲服務平台的相關資料，檢索增強生成模型無法從提供的參考文檔中找到答案。

並在模型中添加提示詞取得檢索文檔與計算檢索增強生成適應性指標，檢驗不同相關性閥值下，使用檢索增強生成適應性指標來分類樣本正負類，假陽性率（FPR）與真陽性率（TPR）的變化。

4.1.1　ROC Curve

本次實驗以 ROC Curve 的方式呈現，結果展示在 Figure 4.1。

doi:10.6342/NTU202401285

Figure 4.1 顯示，在相似度閥值逐步提升的過程中，真陽性率和假陽性率的變化。紅色曲線為該次實驗的 ROC（Receiver Operating Characteristic）曲線，藍色直線為隨機猜測的基準線。每一個點代表特定相似度閥值下的真陽性率（True Positive Rate）與假陽性率（False Positive Rate），相似度閥值從 0.8 逐步提升至 0.99。

除此之外，該次實驗的 AUC（Area Under the Curve）值為 0.8836。

在 Figure 4.1 中，紅色曲線明顯高於藍色隨機猜測的基準線，而實驗中實際紀錄的 AUC 值趨近於 1，表明在檢索增強生成適應性指標的協助下，能夠良好地辨別出負類樣本。曲線在圖的左上角快速提升，表示使用檢索增強生成適應性指標來分類樣本的正負類別在低假陽性率下就能達到較高的真陽性率。

在 Figure 4.1 中，相似度閥值從 0.68 開始逐步提升。在相似度閥值為 0.69 以下時，假陽性率與真陽性率均為 0，且沒有變化，而在相似度閥值達到 0.72 時開始快速提升。這表明在低相似度（0.75 以下）閥值下，假陽性率較低，但真陽性率也較低，誤判的內容較少，但召回率（Recall）也相對較低。

隨著相似度閥值的增加，模型的真陽性率和假陽性率均有所提高，但真陽性率的提高速度明顯快於假陽性率，在高相似度閥值（0.85 以上）的情境下，真陽性率較高，可以識別出較多的正樣本，但同時也有較高的假陽性率。

4.1.2 Analysis of F1-Score and F2-Score

在分析採用檢索增強適應性指標做為標記正負類樣本的實驗中，採用 SQuAD 資料集中文章片段（Context）做為檢索文檔，與 SQuAD 資料集中中根據文章片段設計的提問（Question）做為該次實驗樣本的提問，並採用 SQuAD 資料集中對於每個提問「是否可以根據文章片段回答」的標記，一共選取了 1485 筆資料做為

doi:10.6342/NTU202401285

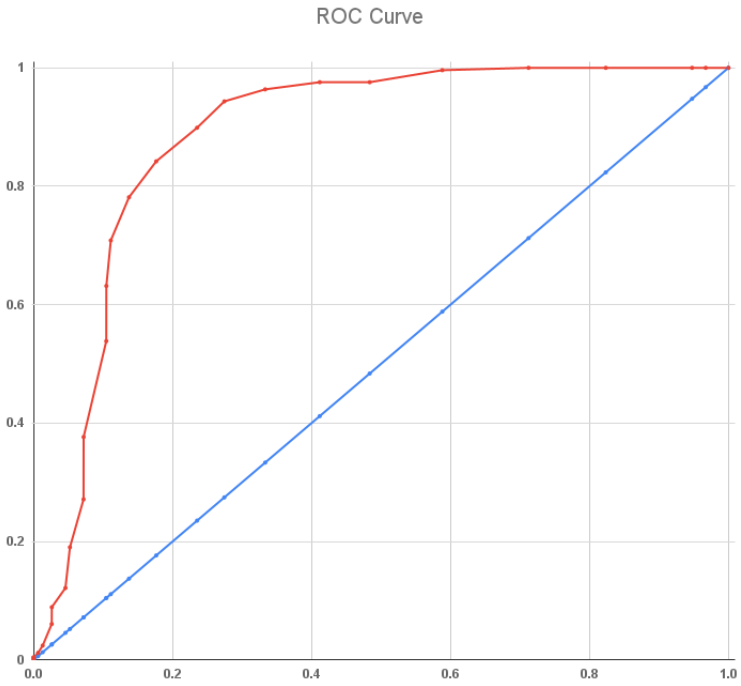

Figure 4.1: 雲服務平台 FAQ 檢索增強生成模型在不同相關性閥值下的特徵曲線（ROC Curve）

樣本，其中 992 筆為「可以從檢索文檔中找到答案」，493 筆為「無法從檢索文檔中找到答案」。

該次實驗樣本中，標記正負類如果採用隨機分佈，精準性為 0.3324，F1-Score 為 0.3994，F2-Score 為 0.4542。

此次實驗目的為驗證使用檢索增強適應性指標來協助執行特定分類任務的表現，以及相關性閥值對於指標能力的影響。

實驗採用「GPT-3.5-Turbo」做為生成器，並測試相關性閥值為”0.7”, ”0.75”, ”0.8”, ”0.85” 四種情況下的表現，除了測試「主要指標（Primary Metric）」，也同時測試「主要指標與提問與檢索文檔相關性指標（Primary Metric with QRD-Relevance Metric）」的表現，觀察其準確率（Accuracy）、精確率（Precision）、召回率（Recall）、F1-Score 以及 F2-Score 的表現。相較於 ROC 曲線的實驗是評估整體性能，F1-Score 與 F2-Score 更貼近實務應用狀況。

doi:10.6342/NTU202401285

Threshold	Accuracy	Precision	Recall	F1-Score	F2-Score
			Primary Metric		
0.70	**0.8343**	**1**	0.5030	0.6694	0.5586
0.75	0.8337	0.9523	0.5274	**0.6789**	0.5791
0.80	0.7414	0.6122	0.6085	0.6104	0.6093
0.85	0.4943	0.3829	**0.8519**	0.5283	**0.6843**
			Primary Metric with QRD-Relevance Metric		
0.70	**0.8350**	**1**	0.5050	**0.6711**	0.5604
0.75	0.7987	0.7383	0.6065	0.6659	0.6290
0.80	0.7111	0.5530	0.6876	0.6130	0.6557
0.85	0.4424	0.3633	**0.9026**	0.5180	**0.6960**

Figure 4.2: 主要指標（Primary Metric）、提問與檢索文檔相關性指標（Primary Metric with QRD-Relevance Metric）在不同相關性閥值（Threshold）下的表現

在 Figure 4.2 中可以觀察到，在較低相關性閥值的情況下，正類樣本很少被判定為生成過程品質不佳，但依然可以召回部分的樣本，這代表部分負類樣本在特定指標中的值會明顯低於正類樣本。並且將檢索增強適應性指標用於識別生成品質不佳的負類樣本，優於原始的檢索增強生成模型。

配合「QRD-Relevance Metric」使用時，指標對於標記負類樣本有更高的敏感度，但誤判的機率也隨之提升，兩者變化的程度相近，代表使用「QRD-Relevance Metric」對整體表現幫助不大。

在此次實驗中，高相關性閥值（0.85），有 85% 的召回率，配合「QRD-Relevance Metric」可以到 90% 的召回率，代表指標可以有效識別大部分生成品質不佳的樣本，F2-Score 也有相對較好的表現，但 F1-Score 的分數卻不太好，整體準確率更是只有 49%，相對於較低的相關性閥值，精確率上的表現在相關性閥值提升時顯著的下降。

為了分析召回樣本的原因，Figure 4.3 中顯示了在不同相關性閥值下，會用於

doi:10.6342/NTU202401285

Threshold	QRD-Relevance	SDRD-Relevance	SDA-Relevance	SDQ-Relevance
0.70	0.2617	0.0325	0.2880	0.2819
0.75	0.3529	0.0365	0.3103	0.3306
0.80	0.6370	0.0629	0.3752	0.4564
0.85	0.7221	0.1258	0.5172	0.6490

Figure 4.3: 檢索增強適應性指標在不同相關性閥值下不同指標各自標示的負類樣本比例

協助標示負類樣本的指標各自標示多少比例的負類樣本，結合 Figure 4.2 的結果可以發現，負類樣本中「SDA-Relevance Metric」、「SDQ-Relevance Metric」，會有較多樣本是低於閥值的，因為在 Figure 4.2 中，低相關性閥值的情況下，正類樣本被誤判的比例不高，精確性還是趨近於 1，但同時還是可以標記出部分負類樣本。

如果同時採用「QRD-Relevance Metric」做為分類樣本正負類的指標，在低相關性閥值下與單純採用主要指標的表現差距不大，代表「QRD-Relevance Metric」與「SDA-Relevance Metric」或「SDQ-Relevance Metric」在負類樣本中是會做出相似反應的。

意即檢索文檔的內容與提問相關性不高、品質不良，且在檢索文檔與支持文檔沒有太大的差異情況下，支持文檔與其他內容的相關性也會反應出異常，代表支持文檔確實可以做為反應檢索過程品質不佳的依據。

如 Figure 1.2 中第二筆樣本的內容，在檢索文檔品質不佳的情況下，支持文檔雖然採用檢索文檔，但生成模型在最後「自行產生回答」，以及 Figure 1.2 第五筆樣本的內容，生成模型做出「拒絕回答」的行為。

在 Figure 4.2 中也可以觀察到，將提問與檢索文檔相關性指標「QRD-Relevance Metric」列入綜合評估的標準，可以提升召回率，但同時精確率和 F1-Score 也會下降，這表明「可以從檢索文檔中找到答案」的正類樣本中，檢索

doi:10.6342/NTU202401285

文檔與提問不一定具有高度相關性，同時 F2-Score 上升的幅度也不明顯，代表雖然有多一些負類樣本被標記出來但模型整體的效能提升不多。

因此，可以根據實務需求調整此指標，如果要強調檢索文檔必須要與提問內容保持一致，就可以採用「QRD-Relevance Metric」指標。

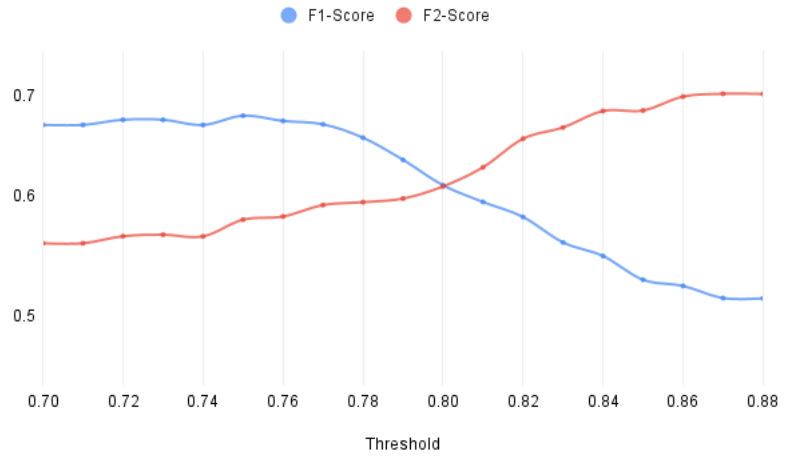

Figure 4.4: 在不同相關性閥值下 F1-Score 與 F2-Score 的表現

Figure 4.4 中，採用 SQuAD 資料集與「GPT-3.5-Turbo」做為生成器，觀察 F1-Score 與 F2-Score 的變化，可以觀察到在本研究使用的資料集中，相關性閥值為 0.8 以下時，檢索增強生成適應性指標更側重於整體的精確度，在 0.8 以上時則更側重於召回率，而 0.8 左右則是代表精確率與召回率是相對平衡的狀態，可以參考這個研究結果與使用需求來制定相關性閥值。

4.1.3 其他生成模型

Figure 4.5 中，顯示選用不同大型語言模型作為檢索增強生成的生成器時，適應性指標的表現。本次實驗採用與前述實驗相同的 SQuAD 資料集，並在替換不同生成模型時依然採用相同的提示詞。

檢索增強生成適應性指標在使用不同的大型語言模型作為生成器時，對於協

doi:10.6342/NTU202401285

Model	Accuracy	Precision	Recall	F1-Score	F2-Score
gpt-3.5-turbo-0613	0.7414	0.6122	0.6085	0.6104	0.6093
gemini-pro	0.6923	0.5308	0.6471	0.5832	0.6199
claude-3-haiku-20240307	**0.7933**	**0.7625**	0.5274	0.6235	0.5620
llama3-8b	0.7347	0.5732	0.6673	0.6167	0.6461
gpt-4o-2024-05-13	0.6681	0.5474	**0.8636**	**0.6701**	**0.7742**

Figure 4.5: 選用不同大型語言模型做為生成器的檢索增強生成適應性指標表現

助識別檢索增強生成過程中品質不佳的負類樣本表現都不錯。越先進的模型如 Claude 3（發表於 2024 年 3 月）、Llama 3（發表於 2024 年 4 月）、GPT-4（發表於 2024 年 5 月）則在不同的面向上有更佳的表現。

在觀察適應性指標與樣本實例後發現，「Claude 3」模型在輸出支持文檔時，更偏好依照檢索文檔的內容生成，較少進行補充、刪減以及做出「拒絕回答」的行為。相較於「Llama 3」，Claude 3 較少以簡答的方式回應，因此在較高相關性閥值的情況下，正類樣本因生成模型產生支持文檔時的調整而被誤判的情況較少。在準確率上有較好的表現，但同時在召回率上的表現則較差。

模型「Llama 3」在綜合表現上良好，實際觀察樣本後發現，「Llama 3」對於負類樣本比較常做出「拒絕回答」的行為，因此在召回率上有不錯的表現。但由於較常以簡答的方式回應提問，使「SDA-Relevance Metric」的平均分數較低，從而影響正類樣本被判定為假陽性的機率，使準確率與精確率較低。

Model	SDRD-Relevance	SDA-Relevance	SDQ-Relevance
gpt-3.5-turbo-0613	0.0629	0.3752	0.4564
gpt-4o-2024-05-13	0.1711	0.4358	0.4827

Figure 4.6: 相同相關性閥值下，「GPT-3.5-Turbo」與「GPT-4o」各項適應性指標對於負類樣本的召回率

doi:10.6342/NTU202401285

相比於「GPT-3.5-Turbo」，模型「GPT-4o」在回答時有較多獨立思考的現象發生。

如 Figure 4.6 所示，GPT-4 在生成支持文檔時，會對檢索文檔進行刪減和總結，並在檢索文檔不足以支持回答時，會修改或自行生成支持文檔的內容。

因此，在相同的相關性閥值下，「SDRD-Relevance Metric」可以召回更高比例的負類樣本。而由於本研究中檢索文檔是根據與提問的相關性所取得的，「SDA-Relevance Metric」在支持文檔變化性更高的情況下，召回率也較「GPT-3.5-Turbo」表現更好。

但也如同 Figure 4.6 所示，「GPT-4o」在輸出回答與支持文檔時，過於積極地刪減和總結內容，雖然對於負類樣本有更好的辨別效果，但有些正類樣本也因為其內容差異性而被標示為負類樣本，造成準確率和精確率的下降。

4.2 輔助判斷指標性能

輔助判斷指標適用於部分情境，Figure 4.7 所示，實驗中採用 SQuAD 資料集，並將相關性閥值設定為 0.8，目的在於實驗追求相對平衡的表現時，各項指標交互使用的表現。

因為 SQuAD 資料集設計的目的是檢驗大型語言模型從文檔片段（Context）中找答案的能力，因此實驗中將文檔片段做為檢索文檔的情況下，「QRD-Relevance Metric」還是有助於提升檢索增強適應性指標分類樣本的表現，而「QA-Relevance Metric」則對於評估檢索增強生成品質沒有顯著幫助，無論是與主要指標一起使用，或是再加上「QRD-Relevance Metric」一同使用，都無法提升整體的分類能力，還會因為回答內容為簡答或延伸回答，造成與提問相關性太低，造成檢索增

doi:10.6342/NTU202401285

強生成適應性指標的整體分類能力下降。

Metrics	Accuracy	Precision	Recall	F1-Score	F2-Score
Primary Metric	**0.7414**	**0.6122**	0.6085	0.6104	0.6093
Primary + QRD	0.7111	0.5530	**0.6876**	**0.6130**	**0.6557**
Primary + QA	0.5279	0.3772	0.6450	0.4760	0.5643
Primary + QRD + QA	0.5253	0.3850	0.7160	0.5007	0.6109

Figure 4.7: 使用主要指標以及主要指標與輔助判斷指標併用時的表現

4.3 區分回答依據

然而，「QA-Relevance Metric」並非沒有用途，在使用雲服務平台 FAQ 資料集為基礎建立的雲服務檢索增強生成模型中，設計了 213 筆大型語言模型具備其知識，但參考文檔中無法取得答題依據的樣本，並以人工方式標示出拒絕回答（Refuse to Answer）的樣本，其中標示出了 39 筆拒絕回答的樣本，拒絕回答的樣本明確表示出了「無法回答」、「檢索文檔中沒有答案」等文字，因此在識別上沒有太大疑慮。

延續上述定義，本次實驗中對樣本標示了「LLM 拒絕回答」的標籤，採用提問與回答的相關性指標（QA-Relevance Metric）閥值 0.8 以下，且主要指標顯示為回答品質不良的樣本，上述條件代表能藉由指標判斷出大型語言模型在檢索文檔品質不良時，產生拒絕回答的行為，如 Figure 1.2 中第二筆資料所示。

此實驗結果顯示，在全部 213 筆樣本中，其中 58 筆被標示為「LLM 拒絕回答」，產生拒絕回答的樣本 39 筆中召回 32 筆，召回率（Recall）為 82%，精確率（Precision）為 55%。

doi:10.6342/NTU202401285

此次實驗顯示提問與回答的相關性指標（QA-Relevance Metric）在協助召回「拒絕回答」的樣本下有不錯的表現，確實可以運用「提問與回答相關性較低」與「檢索增強生成過程不良」同時發生時，來識別「拒絕回答」的現象是否發生。

doi:10.6342/NTU202401285

第五章　Discussion

5.1　實驗結果與應用

如 Figure 4.1 ROC Curve 所呈現，在檢索增強生成適應性指標的協助下，對於標記檢索增強生成模型中生成過程的品質有良好的表現，並在與實際示例的對照中，驗證其具備可解釋性與可追蹤性。

Figure 4.2 中顯示了檢索增強生成適應性指標在不同相關性閥值下可表現出不同的性能特性，能協助完成特定任務。較低的相關性閥值能確保整體結果的正確性，但對於召回生成過程品質不良的樣本表現還有待加強，單獨使用這個相關性低閥值的指標可以改善整體檢索增強生成的表現，但對於消彌生成過程品質不良的現象效果還是有限，而高相關性閥值下，雖然可以召回接近 90% 生成過程品質不良的樣本，但精確度也急遽下降。

召回率以及精確率的平衡需要依使用情境調整，需考量未被召回的樣本是否會危害應用的品質，以及低精確度下再次檢驗所需負擔的成本，因為不同相關性閥值會造成指標性能的偏差，在實務運用上是否能滿足檢索增強生成應用的需求仍有待商榷。

檢索增強適應性指標因為具有相關性分數以及正負類標記的特性，或許可以做為協助調適檢索增強生成架構或微調（Fine-tuning）生成模型的工具。

doi:10.6342/NTU202401285

例如檢驗模型的生成過程的上下文是否可以更趨於一致性，或檢驗修改提示詞後的生成結果，以及利用微調讓生成模型更常做出「拒絕回答」或避免「自行生成回答」的情況發生。

5.2 大型語言模型性能影響

在研究中觀察到，大型語言模型的性能也會影響到檢索增強生成適應性指標的表現。不同的大型語言模型特性同時會反應在適應性指標中。例如，較有信心做出拒絕回答的大型語言模型，對於品質不良的生成過程有較高的召回率；而較能遵循提示詞輸出支持文檔的模型，以及偏好輸出較多回答內容的模型，則對於正類樣本有較少的誤判機率。

也因為檢索增強生成適應性指標能作為分析生成模型特性的工具，模型建構者或許可以通過指標找到符合需求的生成模型。

從研究中可以發現，更先進的大型語言模型由於其認知以及知識的提升，可以識別出更多支持文檔中的明顯錯誤或屬於噪音的內容，並更積極地做出「拒絕回答」、「自行生成回答」或「總結檢索文檔內容」的行為。因此，在本研究實驗中，更先進的大型語言模型的檢索增強生成適應性指標的分數差異更顯著，正負類樣本的區分能力也更強。

同時，大型語言模型生成器對於提示詞的理解和反應可能有所不同，例如，輸出支持文檔的提示詞，大部分情況下會在檢索品質良好時遵照檢索文檔生成，但偶爾也會忽略檢索文檔的情況。或許通過調整提示詞內容，可以提升生成器對於檢索文檔的注意力（Attention）。也可以藉由檢索增強生成適應性指標觀察不同提示詞對於檢索增強生成模型整體效能的影響。

doi:10.6342/NTU202401285

第六章　Conclusion

檢索增強生成適應性指標提供了一個實作範本，包含兩個主要部分：「利用提示詞取得生成模型的認知」和「藉由檢索增強生成中各項內容的相關性來解讀生成過程」。其中，提示詞的設計、取得生成模型認知以及取得檢索增強生成模型內容間相關性的方法都保有調整的彈性，並具有優化潛力，可以根據需求調整或替換為其他的方法。

本研究也證明，相較於原始的檢索增強生成模型，通過取得支持文檔並加入適應性指標，可以更具體地追蹤和解釋檢索增強生成的生成過程。利用適應性指標來協助特定任務，也能取得更好的表現。然而，要完全消弭模型幻覺或追求完全精準的任務表現，仍有提升的空間。

檢索增強生成適應性指標因為具有相關性分數與正負類標記的特性，可以做為訓練、微調模型的依據，並且通過指標可以讓模型開發者分析提示詞以及生成模型對於整體檢索增強生成模型的影響。

如同先前的研究，本研究證明了利用提示詞從生成模型中提取認知和思考過程的意義。即使大型語言模型的效能不斷提升，生成回答的決策過程和模型幻覺問題仍是值得深入探討的課題。

在未來的工作中，可以嘗試使用更先進的生成模型。實驗過程顯示，生成模型的效能與適應性指標的能力成正相關。也可以嘗試結合其他檢索增強生成架構，

doi:10.6342/NTU202401285

讓檢索增強適應性指標協助提升檢索增強生成的過程，從而使整體生成架構能達到更好的表現。

6.1 Limitation

6.1.1 相關性分數與閥值的限制

由於相關性分數的設計是藉由詞嵌入後的向量計算相似度，在模型幻覺產生時，似是而非的文字在相關性的分數表現有可能會比正確回答還要高，實驗中也嘗試將相關性閥值調高到一個極高的分數，但依然有部分負類樣本無法被識別出來，從 Figure 4.2 中也能發現，負類樣本的 F2-Score 在特定閥值後就不再提升。

以樣本內容來看，部分負類樣本的單看文字內容的相關性表現確實與正類樣本差別不大。

在特定專業領域，例如法律、醫學等，些微的字句差異在語意上就有很大的影響，單純以文字的相關性判別出來的正類樣本也可能包含模型幻覺產生出來的內容，這是以詞嵌入計算相關性的方法無法處理的。

並且，雖然在研究中有觀察到，將不同指標套用不同的相關性閥值，會影響到最後辨別正負類樣本的能力，但因為研究過程中選用資料集的特性，無法賦予其解釋性以及驗證其正確性，即指標套用不同閥值能與資料集中哪些標記內容相符，這個議題可以做為日後研究的內容。

doi:10.6342/NTU202401285

6.1.2 僅適用於大型語言模型生成器

在 Self-RAG 或 ReAct 等模型中，更側重於整體檢索增強生成流程的設計，檢索的資料已經不再是單純的利用相似度來找文檔，可能會通過更複雜的方式來完成檢索過程。例如，檢索過程可能會利用查詢語法與資料庫互動、使用關鍵字來找到搜索結果，或者通過多模態輸入（如圖像、音頻等）來輔助生成。在上述情況中，本研究提出的檢索生成適應性指標不一定能適用於所有任務。

本研究所提出的檢索增強生成適應性指標主要基於文字資料進行詞嵌入處理，並通過提示詞提取支持文檔，因此其適用範圍有限。目前，這些指標僅適用於以大型語言模型為生成器的檢索增強生成模型，且不支援多模態輸入的內容。在這些特定應用中，適應性指標能夠有效評估模型的生成質量和可靠性，但在涉及其他數據類型或複雜檢索流程的應用中，這些指標的適用性和有效性可能會受到限制。

雖然本研究的檢索增強生成適應性指標在特定應用中表現良好，但其適用範圍有限，未來需要進一步的研究來擴展其應用場景。

6.1.3 實驗過程限制

本研究實驗設計是藉由檢索文檔品質不良來代表生成品質不良的樣本，並藉由檢索增強生成適應性指標來辨別出這些樣本。

實際情況下有檢索品質良好但生成器產生幻覺的樣本發生，雖然數量不多，但這是實驗中未討論的部分，此情境生成模型產生具有高度相關性的文字內容，但實際無法回答提問，在檢驗模型產生的內容品質上，客觀的分數設計是一個困難且值得研究的課題。

doi:10.6342/NTU202401285

未來或許可以找到評斷這類型樣本的方法，協助檢驗檢索增強生成適應性指標對於此類樣本的衡量能力。

以檢索文檔支持回應與否的資料集來做適應性指標的性能評估有一定的參考價值，但實驗內容還是側重於「生成器」的表現，因為支持文檔是依賴於生成器所產生，所以實驗內容以生成器為主，並以檢索器的檢索結果來判斷檢索品質，在檢索過程上關注較少，另外，先前的研究中已有許多研究在關注檢索過程的表現。

本實驗中提供檢索的資料品質較好，缺少混淆性強的內容，因此在實驗過程中，檢索文檔品質的好壞有較顯著的分數區別，如果能更細部的實驗出適應性指標對於檢索文檔品質判斷的能力，則有助於評估適應性指標中提問與檢索文檔相關性指標，對於不同類型資料集的能力表現。

6.2 Future work

6.2.1 探討提示詞的影響

本研究採用提示詞來取得支持文檔，在不同語言模型中提示詞皆可以提取出支持文檔，但提示詞的設計是否會影響支持文檔的內容是本研究中沒有實驗到的，根據先前研究表示，適當的提示詞修改可以引導或提升大型語言模型特定能力，同時提升檢索增強生成適應性指標藉由支持文檔辨別大型語言模型思維邏輯的表現。

而較先進的大型語言模型可能需要更詳細的提示詞來規範其行為以及提升能力，例如限制 GPT-4o 在生成支持文檔時，過度修改來自於檢索文檔的內容，或是提示 Claude3 在檢索文檔不足以支持回答時，可以自行產生支持文檔的內容。

doi:10.6342/NTU202401285

另外，越聰明的模型在個別工作上的表現也不同，例如提供更細緻的提示詞，「如果認為檢索文檔有問題，那...」，來引導大型語言模型更積極否定不佳檢索文檔內容，自行生成支持文檔，則指標在標記錯誤樣本的表現應該會更好。

因此，在未來的研究中，探究如何藉由提示詞引導生成模型產生良好的支持文檔，以及探討生成檢索文檔與生成回答其中的關係，提升檢索增強生成適應性指標的效能，是值得嘗試的課題。

6.2.2 與其他檢索增強生成架構互動的能力

本研究是在檢索增強生成原有架構中增加內容，對原有架構影響較少，但實驗過程中沒有探究提示詞的添加會對生成模型的回應產生多少影響，如果是依賴於提示詞生成架構可能會有所影響，因此未來可以探究與 Self-RAG、ReAct 等檢索增強生成架構進行互動時，對於整體模型能力帶來的影響，或是可以將檢索增強生成適應性用於改善架構中的哪些部分。

未來的研究可以探討如何擴展這些指標的適用範圍，並研究如何將適應性指標應用於多模態檢索增強生成模型，或探索在更複雜的檢索流程中如何有效地評估生成質量。這可能涉及到開發新的評估方法，或對現有指標進行提升與調整，以適應更廣泛的應用場景。

6.2.3 相關性閾值與指標權重的調整

本研究中，相關性閾值會同時套用至所有指標，即三個指標（QRD-Relevance Metric、SDRD-Relevance Metric、SDA-Relevance Metric）都會使用相同的相關性閾值。然而，未來的研究可以探討各個指標適用的相關性閾值標準，並在不同情境下賦予指標不同的權重值，從而釐清各種情境下，相關性閾值對於整體指標呈

doi:10.6342/NTU202401285

現內容的交互影響。

研究中發現，支持文檔與檢索文檔的相關性指標（SDRD-Relevance Metric）在召回錯誤樣本的比例上略低於其他兩項指標。這可能是因為檢索文檔內容相較於其他內容更容易與支持文檔具有較高的相關性。在制定標準時，「SDRD-Relevance Metric」的閾值應該高於其他指標。這種考量是未來研究可以嘗試的課題，並且可以通過調整生成模型的提示詞來進一步改進。

未來的研究可以更深入地探討個別指標適用的相關性閾值標準，並賦予不同情境下的指標權重值，以提高整體評估的準確性和有效性。

doi:10.6342/NTU202401285

參考文獻

[1] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023.

[2] J. Chen, H. Lin, X. Han, and L. Sun. Benchmarking large language models in retrieval-augmented generation, 2023.

[3] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert. Ragas: Automated evaluation of retrieval augmented generation, 2023.

[4] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang. Retrieval-augmented generation for large language models: A survey, 2023.

[5] Z. JI, N. LEE, R. FRIESKE, T. YU, D. SU, Y. XU, E. ISHII, Y. BANG, W. DAI, A. MADOTTO, , and P. FUNG. Survey of hallucination in natural language generation, 2022.

[6] M. Lee. A mathematical investigation of hallucination and creativity in gpt models, 2023.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, M. L. Heinrich Küttler, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2020.

doi:10.6342/NTU202401285

[8] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, and N. McAleese. Teaching language models to support answers with verified quotes, 2022.

[9] A. Neelakantan, T. Xu, R. Puri, A. Radford, J. M. Han, J. Tworek, Q. Yuan, N. Tezak, J. W. Kim, C. Hallacy, J. Heidecke, P. Shyam, B. Power, T. E. Nekoul, G. Sastry, G. Krueger, D. Schnurr, F. P. Such, K. Hsu, M. Thompson, T. Khan, T. Sherbakov, J. Jang, P. Welinder, and L. Weng. Text and code embeddings by contrastive pre-training, 2022.

[10] Qdrant. Cloud platforms faq dataset.

[11] P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for squad, 2018.

[12] S. J. Semnani, H. C. Z. Violet Z. Yao, and M. S. Lam. Wikichat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia, 2023.

[13] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent abilities of large language models, 2022.

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022.

[15] O. Weller, M. Marone, N. Weir, D. Lawrie, D. Khashabi, and B. V. Durme. "According to . . . " Prompting Language Models Improves Quoting from Pre-Training Data, 2023.

doi:10.6342/NTU202401285

[16] S. Wiegreffe, J. Hessel, S. Swayamdipta, M. Riedl, and Y. Choi. Reframing human-ai collaboration for generating free-text explanations, 2021.

[17] T. Zhao, M. Wei, J. S. Preston, and H. Poon. Automatic calibration and error correction for generative large language models via pareto optimal self-supervision, 2023.

doi:10.6342/NTU202401285